

INTEGRATIVE GENOME-WIDE ANALYSIS OF ALTERNATIVE PRE-MRNA
SPLICING REGULATION BY THE *DROSOPHILA* SR PROTEIN FAMILY

BY

Todd Christopher Bradley

Submitted to the graduate degree program in Pathology and Laboratory Medicine and
the Graduate Faculty of the University of Kansas in partial fulfillment of the
requirements for the degree of Doctor of Philosophy

Chairperson: Dr. Marco Blanchette

Co-Chair: Dr. Patrick E. Fields

Dr. Joan Conaway

Dr. T. Rajendra Kumar

Dr. Julia Zeitlinger

Date Defended: January 16th, 2013

The Dissertation Committee for Todd Christopher Bradley certifies that this is the
approved version of the following dissertation:

INTEGRATIVE GENOME-WIDE ANALYSIS OF ALTERNATIVE PRE-MRNA
SPLICING REGULATION BY THE *DROSOPHILA* SR PROTEIN FAMILY

Chairperson: Dr. Marco Blanchette

Co-Chair: Dr. Patrick E. Fields

Date approved: January 16th, 2013

Abstract

Alternative pre-mRNA splicing is a powerful mechanism that is exploited by higher eukaryotes to diversify their proteomes, and to differentially regulate the expression, function, and localization of mRNA and proteins. Pre-mRNA splicing is typically regulated by RNA-binding proteins that recognize *cis*-acting RNA elements, and either activate or repress splicing of adjacent exons in a temporal, and tissue specific, manner. Understanding how RNA-binding proteins control the splicing code is fundamental to understanding organismal development and disease.

The SR proteins are a well-conserved class of RNA-binding proteins that have an essential role in the regulation of splice site selection, and have also been implicated as key regulators during other stages of RNA metabolism. The complexity of the RNA targets, and specificity of RNA binding location remains poorly understood for many members of the SR protein family. Here, we present a comprehensive study to elucidate how the SR proteins coordinate to regulate alternative pre-mRNA splicing (AS) in *Drosophila*.

Genome-wide analysis of SR-dependent splicing by RNA-seq, reveals that SR proteins are required for the regulation of many types of alternative splicing events, and can act as positive or negative regulators of splice site choice depending on their binding location on the target RNA. In addition, a vast majority of regulated targets require multiple SR protein members for regulation. RNAi of multiple SR proteins simultaneously results in an additive change in the magnitude of splicing. This indicates that SR proteins co-

regulate alternative splicing events in a combinatorial manner through binding specific locations on the target transcripts.

Using single-nucleotide resolution UV cross-linking and immunoprecipitation (iCLIP), we define the specific RNA-binding maps of the entire family of SR proteins in the transcriptome of *Drosophila* S2 cells. We find that SR proteins bind a distinct, but functionally diverse, class of RNAs that includes mRNAs, both constitutive and alternatively spliced, as well as non-coding RNAs. Closer analysis of the bound transcripts revealed that while individual SR proteins can bind unique transcripts, multiple SR protein family members bind a majority of the target transcripts.

This comprehensive analysis reveals position-dependent RNA splicing maps, *in vivo* consensus binding motifs, and a high level of cross- and coordinated regulation of alternative splicing by the SR protein family.

Dedication

To my grandmother, Stella, who taught me the importance of observation.

Acknowledgements

To my mentor, Dr. Marco Blanchette, I cannot fully express my gratitude for the opportunities that you have given me throughout my graduate study. Your enthusiasm and broad scientific knowledge have inspired me throughout my studies. You have created a laboratory that is intellectually challenging, that pushes the frontiers of science by utilizing the latest advances in technology. You gave me the freedom to pursue my research goals, and were there to mentor me when I went astray. You are an exceptional and committed scientist, and I hope to follow in your footsteps.

To my previous undergraduate advisor Dr. Dean Stetler, it is because of you, I am earning a Ph.D. You convinced a stubborn, premed undergraduate student, determined to go to medical school, that advances in healthcare start with discovery. It was in your lab that I fell in love with science. Your encouragement and support, during, and after my time in your lab were instrumental in the pursuit of my Ph.D.

To Nehemiah Alvarez and Scot Harms, thank you for your friendship, and bringing comedy to the lab. You were always willing to assist me with my project, and were a welcome comfort during the frustrating parts of doing research. I will never forget our discussion of the world's problems, and our many potential business ventures. Also, thank you for your constructive criticism of my work, all of my studies have benefited from your contributions.

To Malcolm Cook, you are a bioinformatics wizard. My project would have been impossible without your help and guidance. You have a special way of calming me down,

and working through the many issues that came up in my project. I will be forever grateful.

To the past and present members of the Blanchette lab, Lynnette Gutchewsky, Theophilus Tettey, Dr. Jennifer Knapp, Dr. Nazmul Haque, Suman Chaudhary, David Rebecchi, thank you for making the laboratory an enjoyable place to work, and for all of your assistance with my project along the way.

To Dr. Peter Baumann and members of his laboratory, specifically, Dr. Jay Sarthy, Dr. Martina Gaspari, Ram Kannan, Dr. Aracely Lutes, and Dr. Wen Tang for consistent input on my project, every step of the way, and for great discussions at RNA club.

To the members of my graduate and comprehensive examination committees, Dr. Joan Conaway, Dr. Patrick Fields, Dr. T. Rajendra Kumar, Dr. Philippe Prochasson, and Dr. Julia Zeitlinger, thank you for bringing new and interesting perspectives to my work, and for all of the constructive criticism regarding my dissertation project. I want to thank you for all of the time and effort you put forth in making sure I put the best scientific product forward. I am a better scientist because of you.

To Ms. Shelly Hornbuckle and Ms. Zoë Baldwin, thank you for all of your assistance in helping me navigate the administrative challenges and paperwork that a graduate student faces.

To the past and present directors of the graduate program in the Department of Pathology and Laboratory Medicine, specifically, Dr. Soumen Paul, Dr. Michael Soares, and Dr. Jay Vivian, thank you for making our department a productive and collaborative environment

for graduate students. Also, thank you for constantly checking on my progress, and offering help anywhere you could.

To the Stowers Institute for Medical Research, thank you for providing one of the best places to do science. The core facilities, especially the molecular biology core, and advice from other members of the institute, were instrumental in giving me the ability to tackle any scientific question, no matter how hard it seemed. I will never forget the environment of collegiality and common purpose the institute fostered.

To my friends and fellow students, thank you for your advice, encouragement and emotional support along the way. To Dr. Raul Diaz, Ashleigh Fritz, Dr. Dave Williams, and Mitch McGill, you helped make my first year in graduate school, and the ones that followed, some of the best ever.

To my friend Tony Wu, thank you for your tremendous support over the years. I could always count on you for a lighthearted conversation, and an obscene amount of fun. I will cherish our friendship always.

To my parents, Chris and Pam, my grandparents, Raymond and Stella, and my brothers, Tyler and Anthony, thank you for telling me to go out into the world and do what I love, and making me feel like I can accomplish anything.

To my fiancé's family, Randy, Laura, Josh, and the rest of the clan, thank you for welcoming me into your family, and for your unfaltering encouragement and support. I would not be where I am today without you.

To Teddy, you don't understand what a Ph.D. is, or even care, and that is what I love about you. Thank you for brightening my day with your wagging tail, and forcing me to go cherish the simple things in life.

To my best friend and fiancé, Elyse, I can't thank you enough for your patience and encouragement over the years. Thank you for putting in countless hours reading my work, and listening to my practice talks. You can be proud to be one of the few non-science individuals that actually knows what alternative pre-mRNA splicing is. I am inspired each and every day by your kindness, dedication, and hard work. I hope to one day be half the force in the world as you are. I love you.

Table of Contents

Abstract.....	iii
Dedication	v
Acknowledgements	vi
Table of Contents	x
List of Tables and Figures.....	xiv
List of Abbreviations	xvii
Chapter I: Introduction.....	1
1.1 Introduction.....	1
1.2 Pre-mRNA Splicing.....	1
1.3 Alternative pre-mRNA splicing of metazoan genes	3
1.4 Alternative splicing and disease.....	5
1.5 Patterns of alternative splicing.....	6
1.6 Regulation of Alternative pre-mRNA Splicing	8
1.7 <i>Trans</i> -acting splicing regulators	10
1.8 The Serine/Arginine-rich family (SR proteins)	11
1.9 Methods for the genome-wide detection of alternative splicing.....	17
1.10 Identification of the genome-wide RNA binding distribution of splicing factors.....	20
1.11 Integrating genome-wide datasets: RNA splicing maps.....	23
1.12 Discussion	26

Chapter II: Identification and Quantification of SR Protein Regulated RNA

Processing Events Using RNA-Seq.....	28
2.1 Introduction.....	28
2.2 Materials and Methods.....	31
2.2.1 Drosophila S2 cell culture.....	31
2.2.2 SR protein RNAi.....	31
2.2.3 Generation of recombinant SR proteins for rabbit immunization	33
2.2.4 Generation of recombinant protein and antibody affinity purification	35
2.2.5 Verifying antibody specificity and SR protein RNAi knockdown	38
2.2.6 Illumina next-generation sequence sample library preparation and sequencing	38
2.2.7 Alternative splicing event analysis (ASEA)	39
2.2.8 RT-PCR validation of affected targets.....	40
2.2.9 Identifying changes in overall transcript abundance	42
2.3 Results.....	42
2.3.1 Identification of genome-wide SR-dependent alternative splicing events	42
2.3.2 SR proteins can function as activators and repressors of splicing events.....	49
2.3.3 SR proteins collaborate to regulate a majority of AS events	53
2.3.4 SR proteins regulate 5' and 3' end processing of target transcripts	59
2.3.5 SR proteins are required for maintaining the cellular transcriptional profile	69
2.4 Discussion	72

Chapter III: SR Protein Domain Contribution to Alternative Splicing

Regulation	79
3.1 Introduction.....	79

3.2 Materials and Methods.....	81
3.2.1 Construction of 2x Flag-HA XL6 WT and Mutant constructs	81
3.2.2 Generation stably transfected cell lines	82
3.2.3 Generation of dsRNA against XL6 5' and 3' UTRs.....	83
3.2.4 Expression of WT and Mutant XL6 constructs in <i>Drosophila</i> S2 cells	83
3.2.5 Western-blot to determine construct expression and endogenous XL6 knock-down.....	84
3.2.6 RT-PCR of SR protein targets	84
3.3 Results.....	84
3.4 Discussion	89
 Chapter IV: Generation of Genome-wide RNA Binding Distribution For The SR Protein Family	 93
4.1 Introduction.....	93
4.2 Materials and Methods.....	96
4.2.1 Construction of 2x Flag-HA SR protein constructs	96
4.2.2 Generation stably transfected cell lines	97
4.2.3 iCLIP protocol	98
4.2.4 High-throughput sequencing and mapping.....	103
4.2.5 Identification of iCLIP-Seq clusters	103
4.2.6 Motif analysis.....	104
4.2.7 RNA splicing maps	104
4.3 Results.....	105

4.3.1 Mapping SR protein binding to RNAs at nucleotide resolution with iCLIP-Seq	105
4.3.2 Global landscape of RNA transcripts bound by SR proteins.....	110
4.3.3 Consensus, <i>in vivo</i> , binding motifs for the SR proteins.....	114
4.3.4 SR protein RNA splicing maps.....	120
4.4 Discussion	125
Chapter V: Perspectives and Conclusions.....	131
5.1 SR proteins and human disease.....	131
5.2 Combinatorial regulation of AS by the SR proteins	132
5.3 AS regulation requires and RS-domain, and specificity is determined by the RRM-domain	134
5.4 Global RNA-binding landscapes of the SR proteins	135
5.5 Final thoughts	137
Chapter VI: References.....	139

List of Tables and Figures

Figure 1.1: Catalytic steps of pre-mRNA splicing.....	2
Figure 1.2: Spliceosome assembly.....	3
Figure 1.3: Schematic representations of AS events important for development and disease	8
Figure 1.4: Common patterns of alternative mRNA processing.....	9
Figure 1.5: <i>Cis</i> -elements in a pre-mRNA important for proper splicing control.....	10
Figure 1.6: SR protein paralogs	12
Figure 1.7: Roles of SR proteins in constitutive and alternative splicing.....	14
Figure 1.8: Uncovering the splicing code	25
Figure 2.1: Experimental strategy for identification of SR protein regulated AS events	44
Figure 2.2: RNAi of the eight members of the <i>Drosophila</i> SR protein family in <i>Drosophila</i> S2 cells.....	45
Figure 2.3: RNA-Seq of RNA extracted from control and SR-depleted S2 cells.....	47
Figure 2.4: SR proteins regulate all types of simple AS events.....	49
Figure 2.5: Gene ontology analysis of SR regulated genes	51
Figure 2.6: Genome-wide analysis of SR-dependent AS events	53
Figure 2.7: Proportion of events that are less included upon depletion (SR-activated, red), or more included upon depletion (SR-repressed, dark gray) by AS type (A-E). 55	
Figure 2.8: Differentially regulated events detected by RT-PCR.....	57
Figure 2.9: Multiple SR proteins co-regulate AS events	59
Figure 2.10: Combinatorial regulation of AS by the SR proteins.....	60

Figure 2.11: RT-PCR demonstrates combinatorial splicing control of specific SR targets.....	63
Figure 2.12: SR proteins regulate transcript start sites	65
Figure 2.13: SR proteins regulate alternative polyadenylation that changes the protein coding sequence.....	67
Figure 2.14: SR proteins regulate alternative polyadenylation in the 3' UTR independent of AS.....	69
Figure 2.15: SR proteins regulate RNA transcript levels	72
Figure 2.16: Over-represented GO-terms in SR protein gene expression regulated targets.....	74
Figure 2.17: SR protein regulation of gene expression is independent of AS regulation	75
Figure 3.1: SR protein domain organization.....	80
Figure 3.2: Diagram of WT XL6 and variant proteins	85
Figure 3.3: Expression of recombinant XL6 constructs in <i>Drosophila</i> S2 cells	86
Figure 3.4: Specificity of SR protein AS regulation depends on RRM-domain, and requires the RS-domain.....	87
Figure 4.1: Schematic overview of iCLIP procedure	99
Figure 4.2: Analysis of cross-linked SR protein-RNA complexes using iCLIP.....	108
Figure 4.3: iCLIP mapping information	109
Figure 4.4: SR proteins bind to distinct positions in <i>Drosophila</i> RNAs	111
Figure 4.5: SR proteins bind a large number of genes.....	113
Figure 4.6: SR proteins bind a similar set of genes at unique positions	116

Figure 4.7: Distribution of SR protein CLIP clusters within RNA classes and transcript regions.....	117
Figure 4.8: SR proteins bind to specific motifs, <i>in vivo</i>	120
Figure 4.9: Distribution of SR protein motifs in regulated AS targets	122
Figure 4.10: SR protein binding near splice sites	123
Figure 4.11: SR proteins bind in a position-specific manner around cassette exon .	125
Figure 4.12: SR proteins bind in a position-specific manner around competing 5' splice sites.	127

List of Abbreviations

3' ss	3' slice site
5' ss	5' splice site
aa	Amino acid
APA	Alternative polyadenylation
APR	Alternative promoter
ARE	AU-rich elements
AS	Alternative Splicing
ASE	Alternative splicing event
ASF/SF2	Splicing factor 2
ASP	Antisense primer
ATP	Adenosine triphosphate
BCL-xL	B-cell lymphoma-extra large
BCL-xS	B-cell lymphoma-extra short
bp	Branch point
bp	Base pair
c-src	non-neuronal SRC tyrosine kinase
cDNA	Complementary DNA
CE	Cassette Exon
CHIP	Chromatin immunoprecipitation
CLIP	Cross-linking immunoprecipitation
Clk/Sty	Cdc2-like kinase
CR-APA	Coding region-alternative polyadenylation
CTD	C-terminal domain
DNA	Deoxyribonucleic acid
dNTP	deoxynucleotide triphosphate
DppIII	Dipeptidyl aminopeptidase III
dsRNA	Double-stranded RNA
EMSA	Electrophoretic mobility shift assay
ESE	Exonic splicing enhancer
ESS	Exonic splicing silencer
Fox-1	Forkhead box-1
Fox-2	Forkhead box-2
GO	Gene ontology
HIV-1	Human immunodeficiency virus-1
hnRNP	heterogeneous nuclear ribonucleoprotein particle
HRP	Horseradish peroxidase
Hsf	Heat shock factor
	Individual nucleotide resolution cross-linking
iCLIP	immunoprecipitation
Indy	I'm not dead yet

ISE	Intronic splicing enhancer
ISS	Intronic splicing silencer
Klar	Klarsicht
miRNA	micro RNA
mj	Milijoules
ml	Milliliter
mM	Milimolar
MNK2	MAPK signal-integrating kinase-2
mRNA	messenger RNA
mTOR	Mammalian target of rapamycin
n-src	neuronal SRC tyrosine kinase
ncRNA	Noncoding RNA
Nfat	NFAT homolog
NFX1	Nuclear RNA export factor-1
ng	Nanograms
NMD	Nonsense-mediated decay
NOVA	Neuro-oncological ventral antigen
nt	Nucleotide
oligos	Oligonucleotides
PAR-CLIP	Photoreactive ribonucleoside-CLIP
PBS	Phosphate buffered saline
PCR	Polymerase chain reaction
Pep	Protein on ecdysone puffs
PNK	Polynucleotide kinase
Pol II	RNA polymerase II
PSI	Percent spliced in
PTB	polypyrimidine tract-binding protein
Pum	Pumillo
Py	polypyrimidine tract
RBP	RNA binding protein
RIP	RNA immunoprecipitation
RNA	Ribonucleic acid
RNA-seq	RNA sequencing
RNAi	RNA interference
RPKM	Reads per kilobase of million mapped reads
RRM	RNA recognition motif
RS-domain	Arginine-serine rich domain
Rsfl	Repressor splicing factor-1
RT-PCR	Reverse-transcription polymerase chain reaction
Sam68	Src-associated in mitosis, 68 kDa
SDS	Sodium dodecyl sulfate
SELEX	Selected evolution of ligands through exponential enrichment
Skd	Skuld

SMA	Spinal muscular atrophy
SMN1	Survival of motor neuron 1
SMN2	Survival of motor neuron 2
snoRNA	small nucleolar RNA
snRNP	Small nuclear ribonucleoprotein particle
SP	Sense primer
SR protein	Serine-arginine rich protein
SRE	Splicing regulatory element
SRPK1	SR-protein kinase-1
SRSF1	SF2/ASF SR protein
SRSF3	SRp20 SR protein
SRSF4	SRp75 SR protein
Su(z)12	Suppressor-of-zeste-12
SV40	Simian vacuolating virus 40
Sxl	Sex lethal
Syb	Synaptobrevin
T7	Bacteriophage T7 promoter
TBST	Tris buffered saline with Tween
Ten-m	Tenascin major
Tra	Transformer
Tra2	Transformer-2
tRNA	Transfer RNA
U2AF	U2 snRNP auxiliary factor
UTR	Untranslated region
UTR-APA	Untranslated region-alternative polyadenylation
UV	Ultraviolet
Vps35	Vacuolar protein sorting 35
μg	Microgram
μl	Microliter
μM	Micromolar

Chapter I: Introduction

1.1 Introduction

After transcription, the removal of non-coding sequences, introns, from the pre-mRNA by pre-mRNA splicing allows the coding sequences, exons, to be spliced together to form functional mature RNAs (mRNAs). This process is an essential step for gene expression in higher eukaryotes. Alternative splicing (AS) is the process of generating multiple mRNA isoforms from a single pre-mRNA by joining exons together in different combinations.

1.2 Pre-mRNA Splicing

Pre-mRNA splicing takes place in two transesterification reactions (Figure 1.1). In the first step, the 2'-hydroxyl group of the adenosine at the branch point attacks the phosphate at the 5' splice site. This results in the detachment of the 5' exon, and an intron-lariat attached to the 3' exon. The next step proceeds with the attack of the 3' hydroxyl group of the 5' exon on the phosphate bond at the 3' splice site, resulting in release of the intron lariat, and ligation of the exons [1].

A macromolecular complex termed the spliceosome carries out the splicing reaction. The spliceosome consists of five small nuclear ribonucleoprotein particle (snRNPs), and hundreds of accessory protein components [2, 3]. The spliceosome recognizes sequence elements within the intron that define the intronic boundary, and promote excision of the intron at precise locations. Assembly of the spliceosome is initiated by recognition of the 5' and 3' splice sites by U1 snRNP and U2AF (U2 snRNP auxiliary factor), respectively

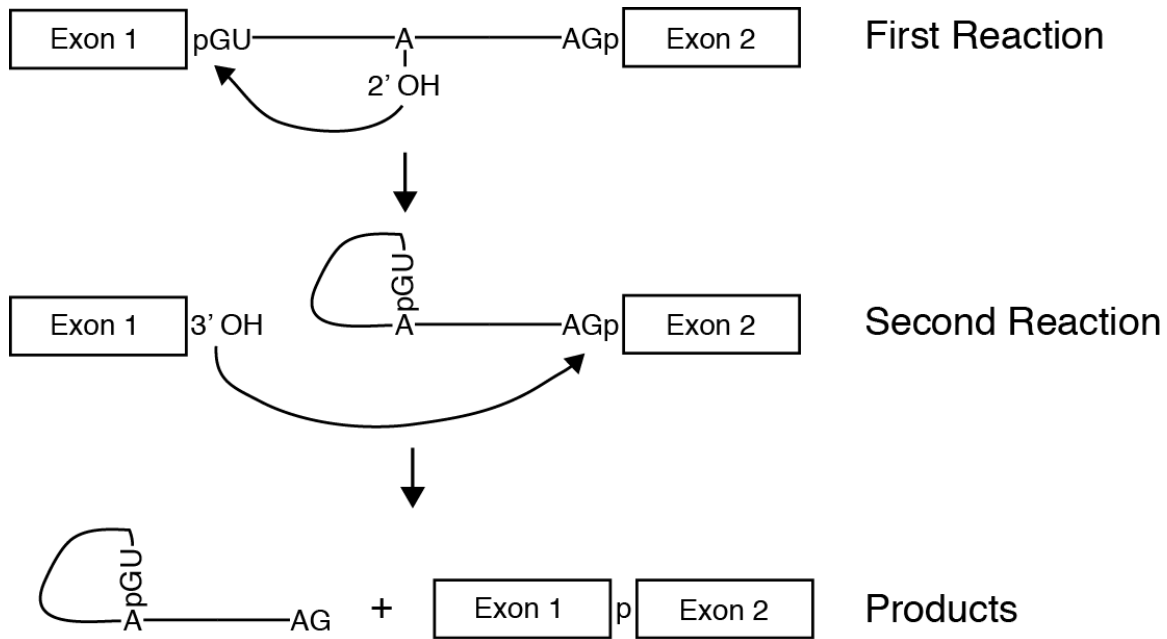


Figure 1.1: Catalytic steps of pre-mRNA splicing. Pre-mRNA splicing is carried out in two transesterification reactions. The first reaction results in the release of the first exon and formation of an intron lariat fused to the second exon. The second reaction produces spliced exons and a liberated intron lariat.

(Figure 1.2). This results in the formation of the E complex. Next, the U2 snRNP is recruited to the branch point sequence, in an ATP-dependent manner, resulting in A complex formation [4]. This leads to the subsequent recruitment of the U4, U5, and U6 tri-snRNP forming the B complex. The B complex undergoes a series of complex rearrangements that allows for the formation of the C complex, and a catalytically active spliceosome. The C complex leads to splicing, release of the intron lariat, and subsequent spliceosomal recycling [5]. Splicing together exons can be either constitutive (when the exon is always a part of the final mRNA transcript), or alternative (when the exon is not always included in the final mRNA).

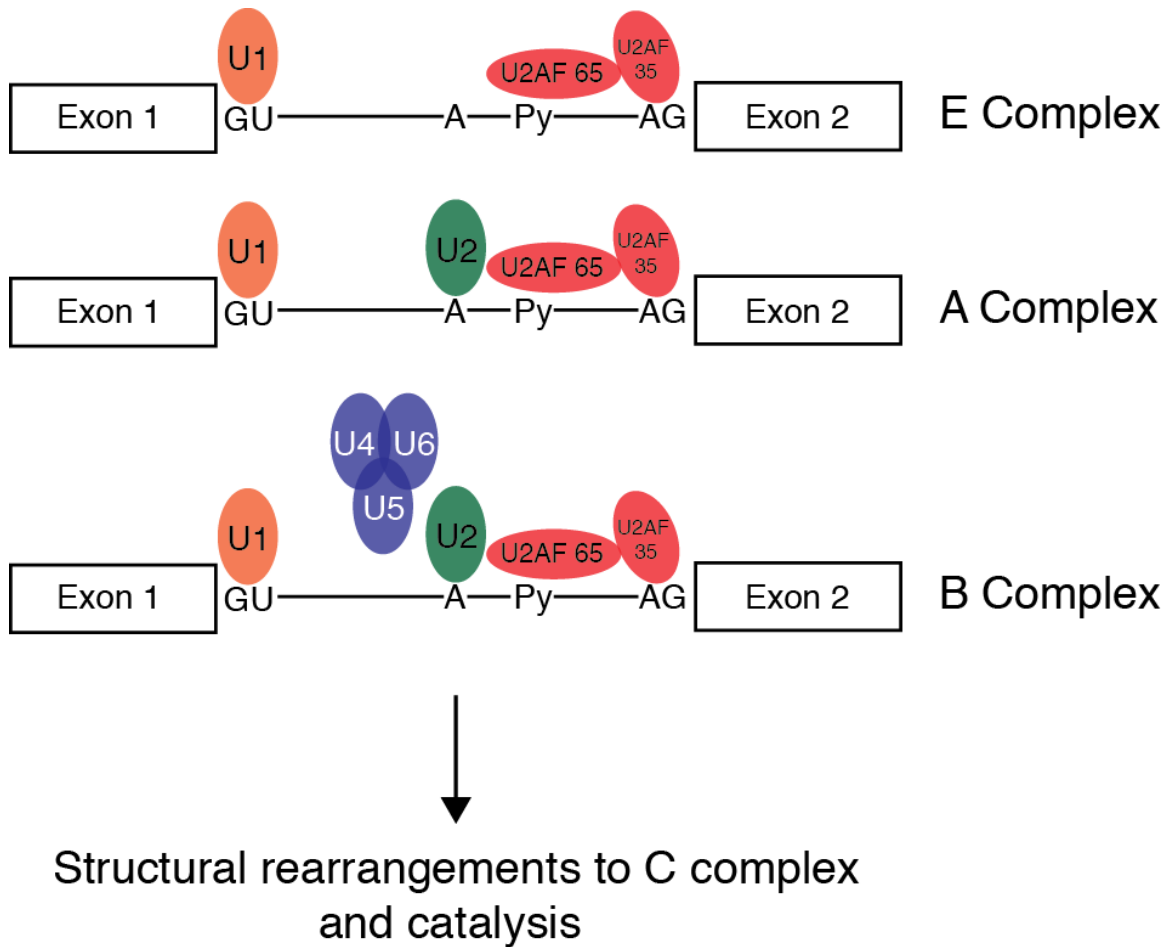


Figure 1.2: Spliceosome assembly. During E complex formation the U1 snRNP and U2AF bind the 5' and 3' splice sites, respectively. Next, the U2 snRNP binds the branch point (A) and forms the A complex. Recruitment of the U4/U6 • U5 tri-snRNP forms the B complex. This leads to a series of structural rearrangements and formation of the catalytically active spliceosome.

1.3 Alternative pre-mRNA splicing of metazoan genes

Recent sequencing of higher eukaryotic genomes revealed that *Drosophila melanogaster* and *Caenorhabditis elegans* have ~14,000 and ~38,000 protein coding genes, respectively, while there are ~51,000 protein coding genes in the human genome [6]. This suggests that there is little correlation between the number of genes in an organism's

genome, and the biological complexity of the organism. It has been proposed that other genomic features, such as alternative splicing, can account for the differences in complexity by expanding the coding capacity of the genome [7]. AS significantly expands the diversity of the proteome by producing multiple mRNA transcripts from a single protein-coding gene [8]. Recent estimates suggest that greater than 90-95% of human multi-exon transcripts can undergo AS [9]. In addition to expanding the proteome, AS is utilized to regulate gene expression, as many of the mRNA variants are expressed differentially between tissues, cell types, and developmental conditions [10].

Over the past decade, instances of AS controlling important physiological pathways have been characterized, one of which is the sex determination pathway in *Drosophila melanogaster* [1, 11, 12]. Sex lethal (Sxl) is an RNA-binding protein that is expressed specifically in female flies, where it represses down-stream splicing patterns of other splicing and transcriptional regulators, that would lead to male development (Figure 1.3A) [13]. Regulation of Sxl, and other splicing regulators control sex determination in *Drosophila*. The idea that AS splicing is an important regulatory mechanism can be extended to other organisms, in mammalian cells, there is a specific isoform of the SRC tyrosine kinase that is activated in neurons (n-src), and repressed in non-neuronal cells (c-src). Splicing control of this event is attributed to neural specific splicing regulators nPTB, Fox-1/2, and generally expressed splicing regulators hnRNP H, hnRNP F, and ASF/SF2 (Figure 1.3B) [14-16]. It has been hypothesized that the different SRC variants have different interacting partners, and therefore, different physiological functions between different tissue types. Given the pervasiveness of AS, and that proper control is essential for tissue development and homeostasis, it is important to understand the

mechanism of AS regulation, especially since misregulation of specific AS events have been shown to directly cause several genetic diseases.

1.4 Alternative splicing and disease

Approximately 15% of mutations that result in genetic disease affect sequence elements important for pre-mRNA splicing [17]. One well-studied example is Spinal muscular atrophy (SMA) in humans. SMA is a severe neurodegenerative disorder that results from lack of survival of motor neuron 1 (SMN1) protein. Survival of motor neuron 2 (SMN2) is a paralogous gene that differs from SMN1 by a single nucleotide change (C>T transition in exon 7), which results in the increased skipping of exon 7, and the increased production of a non-functional protein (Figure 1.3C). It is believed that this nucleotide change results in either the disruption of an exonic splicing enhancer, or the creation of an exonic splicing silencer [18, 19]. Several therapeutic approaches are aimed at inducing higher inclusion of exon 7 of SMN2 to produce more functional SMN protein in affected patients [20, 21].

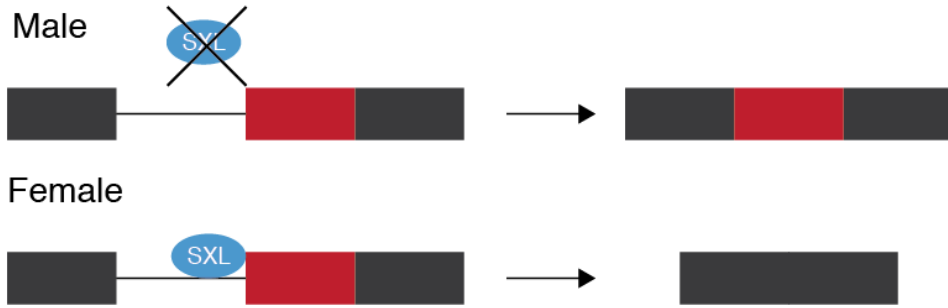
Additionally, mutations that affect the AS of tumor suppressors, leading to their inactivation, are responsible for a variety of inherited and sporadic cancers. One of the earliest discovered examples of AS regulation in cancer was the regulation of Bcl-X (Figure 1.3D). Bcl-X pre-mRNA can be alternatively spliced to produce a long isoform (Bcl-X (L)), which has anti-apoptotic effects, and a short isoform (Bcl-X(S)) which promotes apoptosis [22]. High Bcl-X L/S ratios are observed in a variety of cancer types, consistent with an important role for Bcl-X (L) in cancer cell survival [23, 24]. Often, aberrant expression of splicing factors in cancer cells contribute to the change in AS

observed in tumor suppressors [25, 26]. For example, it was observed that Sam68 overexpression resulted in an increase in Bcl-X(S) isoform, and had a pro-apoptotic effect [27]. Therefore, understanding the physiological targets, and the molecular mechanism of action of splicing regulators, will be important for development of novel strategies to treat splicing-associated disease.

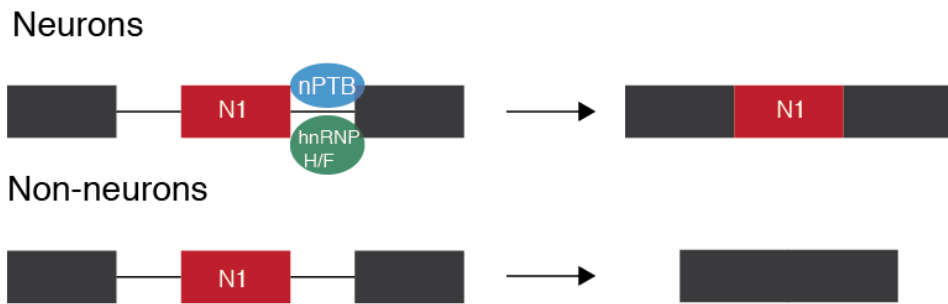
1.5 Patterns of alternative splicing

Common patterns of AS have been observed and are used to classify alternative splicing into five simple types (Figure 1.4) [1, 28]. The most commonly studied forms of AS are cassette exon splicing events. This occurs when an entire cassette exon is included or skipped in the resulting mRNA. There are also competing 5' splice sites and 3' splice sites, where only a portion of the exon is included or skipped. Mutually exclusive exon AS occurs when two or more adjacent cassette exons are spliced such that only one exon in the group is included at a time. The most frequent form of AS observed in *D. melanogaster* are intron-retention events (Figure 2.4B). This is when the entire intron is removed or retained. This form of AS is less common in mammals, perhaps due to the longer length of introns that would significantly disrupt the protein coding sequence if retained. Finally, there are alternative first exons and alternative last exons. These events are uniquely regulated, because they do not solely rely on changes in splice site usage. Alternative transcriptional promoter usage, or alternative transcription termination and polyadenylation, influence the outcome of these events. A single pre-mRNA can exhibit multiple sites of alternative splicing using these simple patterns, in a combinatorial manner, to produce multiple mRNA isoforms.

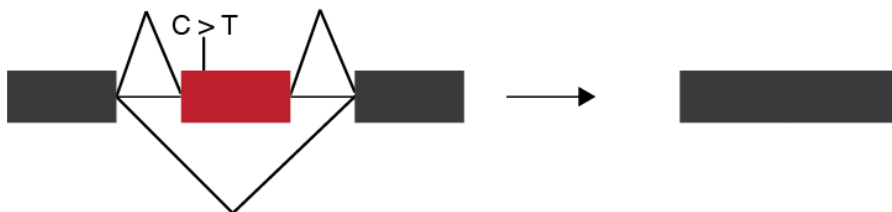
A. Transformer



B. SRC Tyrosine Kinase



C. SMN2



D. BCL-X

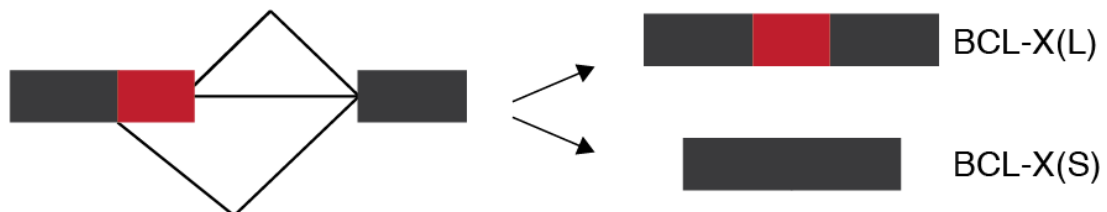


Figure 1.3: Schematic representations of AS events important for development and disease. (A) SXL (blue oval) is a female-specific splicing factor that regulates the splicing of targets in the *Drosophila* sex determination cascade that contribute to female attributes, such as, Transformer. (B) SRC tyrosine kinase has a neuronal expressed isoform that is regulated by the splicing regulators nPTB (blue oval), hnRNP H/F (green oval) and others. (C) SMN2 has a single nucleotide polymorphism that results in the skipping of exon 7, and results in nonfunctional SMN protein. This splicing event is a therapeutic target for SMA patients hoping to restore SMN protein production. (D) BCL-X AS produces a long isoform (BCL-X (L)) that is anti-apoptotic, and a short isoform (BCL-X (S)) that is pro-apoptotic. Several tumors preferentially express the long isoform and evade apoptosis.

1.6 Regulation of Alternative pre-mRNA Splicing

The sequences of the pre-mRNA elements recognized by the splicing machinery to catalyze intron removal are highly degenerate, and therefore occur frequently within the genome, despite being ignored by the spliceosome (Figure 1.5). The sequences of the splice sites within the intron can be classified as “strong” or “weak” depending on their similarities to the consensus splice site motifs, which reflect the strength of recognition by components of the spliceosome. Introns with strong splice sites are efficiently recognized by the spliceosome, and constitutively spliced. Introns with weak splice sites are often alternatively spliced, or require additional regulatory signals for proper splicing [29]. Therefore, in addition to splice sites, other splicing regulatory elements (SREs) within the pre-mRNA are required for proper splicing efficiency, and control of AS (Figure 1.5) [30, 31]. Exonic splicing enhancers (ESEs) and intronic splicing enhancers (ISEs), are bound by RNA-binding proteins that subsequently recruit, and/or promote, the assembly/stability of spliceosomal components, which consequently activate the use of nearby splice sites. Similarly, exonic and intronic splicing silencers (ESSs and ISSs, respectively) are bound by factors that inhibit the use of nearby splice sites [1, 32]. The

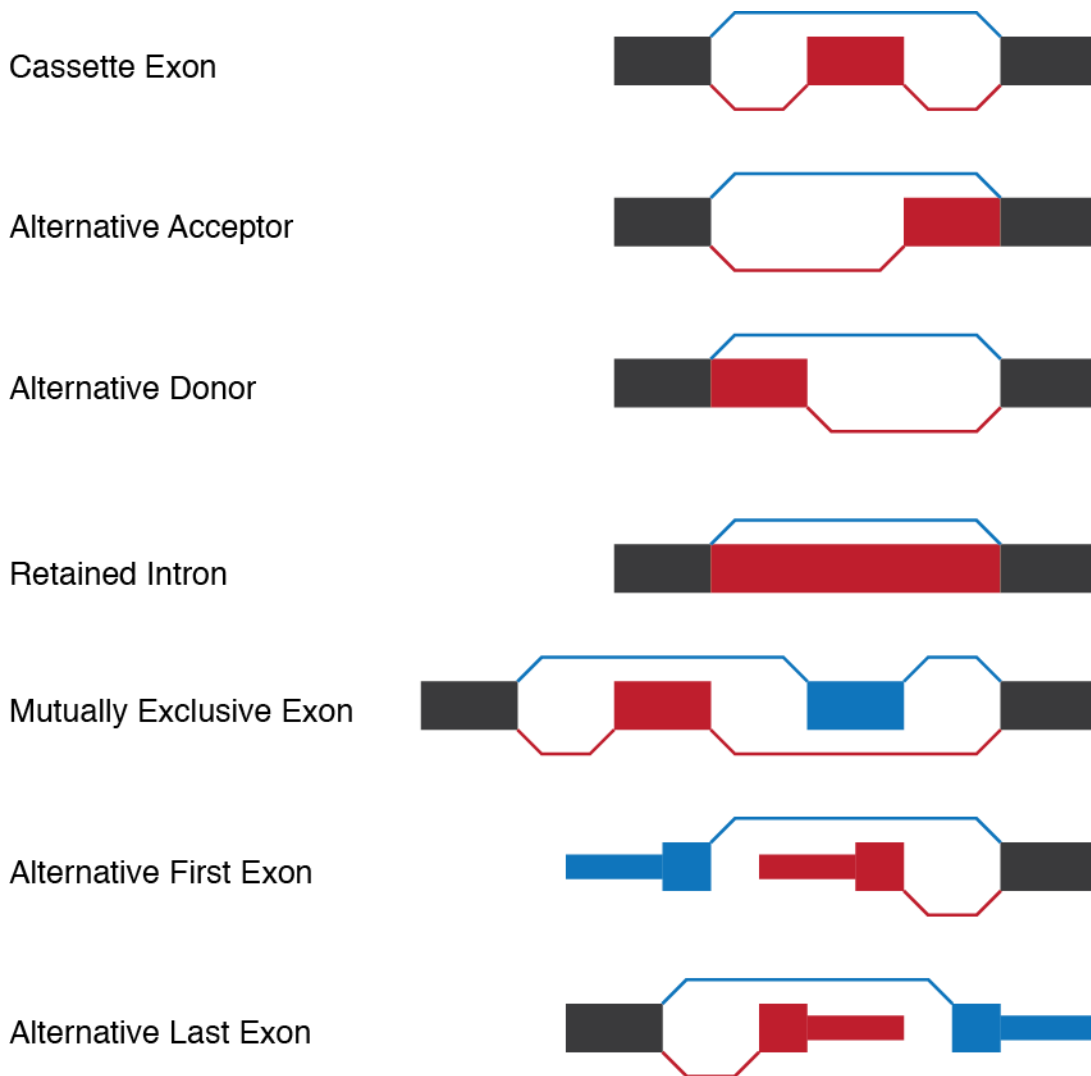


Figure 1.4: Common patterns of alternative mRNA processing. Black boxes represent constitutive exons. Red and blue boxes indicate alternative exons. Thinner portions of alternative exons indicate untranslated regions (UTRs). Red and blue lines correspond to the two outcomes alternative splicing.

specific combinations of these elements, their proximity to splice sites in conjunction with the specific trans-acting factors expression levels, determine which splice sites are used, and contribute to the proper splicing outcome. Besides the splice site consensus



Figure 1.5: *Cis*-elements in a pre-mRNA important for proper splicing control.

Exons are represented as rectangle boxes. 5' and 3' splice site dinucleotides represent the intron boundary (GU-AG, respectively). The branch point, A, and the polypyrimidine tract, Py, sequences are also crucial for splice site recognition. Exonic and Intronic splicing enhancers (ESE, ISE) and silencers (ESS, ISS) are important splicing regulatory sequences that can vary in length and position on the pre-mRNA.

sequences and SREs, RNA secondary structure, RNA polymerase processivity and rate, and epigenetic chromatin marks can influence splice site selection [33, 34].

1.7 *Trans*-acting splicing regulators

There are many groups of splicing regulatory proteins that bind to *cis*-elements in the pre-mRNA to control splice site selection. One class of RNA-binding proteins are proteins that contain arginine-serine rich (RS) domains. Members of the serine-arginine rich (SR) protein family are in this group, and are thought to be general activators of splice site selection by binding to enhancer sequences in the pre-mRNA (ESEs and ISEs) [35].

Another well-conserved class of splicing regulators is the hnRNP family. Members of the heterogeneous nuclear ribonucleoprotein (hnRNP) family do not have an RS-domain, and are generally characterized as repressors of splicing by binding silencing elements in the pre-mRNA (ESSs and ISSs) that block spliceosome assembly, or antagonize splicing activators, such as SR proteins [36]. In addition to the well-characterized SR and hnRNP protein families, there are a number of unrelated splicing regulators that function in a tissue specific manner, such as the RNA-binding proteins NOVA and FOX [37]. The

regulation of AS by many proteins remains poorly understood, as are the contributions of other protein families that do not bind RNA, such as chromatin remodelers.

1.8 The Serine/Arginine-rich family (SR proteins)

The SR protein family is an essential, highly conserved, class of splicing regulators. SR proteins are characterized as having one or two N-terminal RNA recognition motifs (RRM), and a C-terminal domain enriched in arginine and serine dipeptides (RS-domain). In humans, this definition has identified 12 canonical SR proteins, and dozens of related SR-like proteins that contain an RS-domain, but lack a defined RRM were found [38]. The number of introns that interrupt eukaryotic genes varies greatly between organisms. Since SR proteins have been associated with regulation of splicing, it was predicted that the number of SR protein family members would increase with the number of intron-containing genes. Indeed this correlation exists; only two SR proteins have been found in *S. pombe*, seven SR proteins in *C. elegans*, and 12 SR proteins in humans. In *Drosophila*, there have been eight SR proteins identified, and they are highly homologous to their mammalian counterparts (Figure 1.6) [39]. SR proteins were first identified as essential components of the splicing reaction, because of their ability to complement splicing deficient cytoplasmic S100 extracts [40, 41]. SR proteins have since been implicated in the regulation of a large number of AS events, and participate in all crucial aspects of RNA metabolism, including mRNA export, localization, translation, and nonsense mediated decay (NMD) [42-45].

To explore SR protein roles in organismal development, animal and insect knockout models were engineered for several members of the SR protein family. Surprisingly, SR

D. melanogaster gene symbol	H. sapiens		Blast E value
	Original name	NCBI gene symbol	
SC35	SC35	SRSF2	2.00E-51
SF2	ASF/SF2	SRSF1	3.00E-83
B52	SRp75	SFRS4	3.00E-83
	SRp55	SFRS6	2.00E-77
SRp54	SRp35	SREK1 (SFRS12)	5.00E-57
Xl6	9G8	SFRS7	3.00E-36
RBP1	SRp20	SFRS3	2.00E-38
RBP1-like	SRp20	SFRS3	3.00E-35
	9G8	SFRS7	2.00E-36
Rsf1	SRp20	SFRS3	2.00E-21
	9G8	SFRS7	9.00E-24

Figure 1.6: SR protein paralogs. *Drosophila* and mammalian SR proteins share extreme sequence similarity and conservation.

protein-null mice for three SR proteins (SC35, ASF/SF2, and SRp20) all showed early embryonic lethality [46]. Previously, it was thought that SR proteins had redundant functions, because of the ability of multiple SR proteins to complement splicing deficient S100 extracts. Since each SR-null mouse shows lethality when only a single member of the SR protein family is knocked-out, this provided evidence that SR proteins are not all functionally redundant, and individual SR proteins have specific roles during development. However, embryonic fibroblasts derived from these mice are viable. This indicates that, individually, SR proteins are not essential for cellular viability, but rather function tissue specifically and are critical for precise developmental decisions. In addition to mammalian development, knockout of the *Drosophila melanogaster* SR protein B52 also had deleterious effects on development, as B52-null mutants do not

develop past the second-instar larval stage [47]. This observation also supports the non-overlapping roles that specific SR proteins have *in vivo*.

To bypass embryonic lethality, conditional knockout mice have been engineered. Conditional deletion of SC35 in the thymus and heart was shown to cause a defect in T-cell maturation, and to induce dilated cardiomyopathy [48]. Similarly, deletion of ASF/SF2 in the heart caused a hyper-contraction phenotype due to defects in post-natal heart remodeling [49]. Both alternative splicing and gene expression profiles are changed in these mutant tissues, and are thought to contribute significantly to the phenotypes observed. Conditional knockout of B52 in *Drosophila* eye tissue resulted in flies with small eyes. This was attributed to B52 regulation of an AS event of a master regulator of eye development, *eyeless*, resulting in a protein isoform that gives rise to small eyes [50].

Additionally, expression of SR proteins is up-regulated in certain human tumors. Tumors in the lung, ovary, breast, kidney, colon, liver, and pancreas all showed elevated SR protein levels [25, 51, 52]. These studies demonstrate the vital role that SR proteins play in embryogenesis and tissue homeostasis, and further stress that the understanding the molecular targets, and mechanism of SR protein regulation, is crucial to our understanding of organ development and disease progression.

SR proteins are thought to be general activators of splice site selection by binding to ESEs in the pre-mRNA (Figure 1.7). Once bound, SR proteins can recruit and/or stabilize

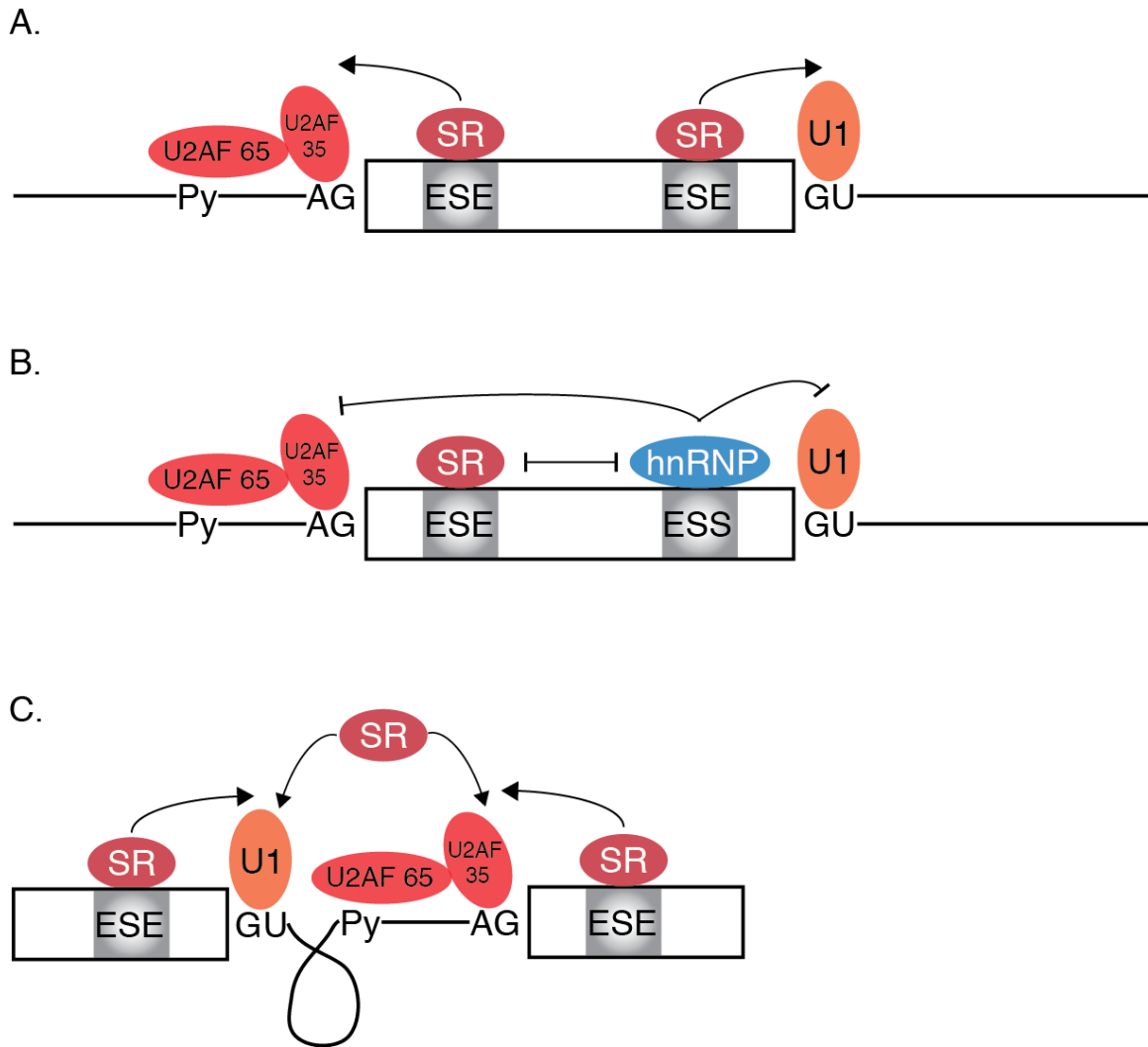


Figure 1.7: Roles of SR proteins in constitutive and alternative splicing. (A) SR proteins bound to ESE elements in the pre-mRNA can recruit and/or stabilize components of the spliceosome (U1 snRNP and U2AF) in a process known as exon definition. (B) SR proteins can antagonize splicing repressors bound to ESS elements, and promote splice site selection. (C) Through RS domain interactions SR proteins can help bridge the intron, and bring the 5' and 3' splice sites in close proximity for more efficient splicing.

components of the splicing machinery, such as the U1 snRNP to the 5' splice site or U2AF to the 3' splice site, through RS-domain protein-protein interactions (Figure 1.7A)

[53-55]. SR proteins can also form a network of protein-protein interactions across introns to juxtapose the 5'ss and 3'ss, to facilitate early spliceosome assembly (Figure 1.7C) [56]. Typically, SR protein binding sites, and binding sites of other splicing regulators, such as the hnRNP family, are in close proximity, suggesting that the interplay between activation and repression modulates the frequency of exon inclusion (Figure 1.7C) [57].

SR protein binding to target sequences in the pre-mRNA are mediated by interactions in the RRM-domain. Using *in vitro* binding studies, there is compelling evidence that individual SR proteins can have different, functionally significant, RNA-binding sites [58]. Also, It has been demonstrated that the position of SR-protein binding sites in the pre-mRNA determines whether these sites will function as enhancers or silencers of exon inclusion [59]. It was also shown that the activation of exon inclusion was proportional to the number of RS dipeptide repeats within the RS-domain of the bound SR protein [60]. In addition, the frequency of exon inclusion for many AS events is determined by the binding position of multiple RNA splicing factors, including multiple SR protein family members. For example, SR proteins 9G8, ASF/SF2, and Rbp1 cooperatively bind with the splicing regulators Tra and Tra2 to specific locations on the *Drosophila doublesex* transcript to regulate sex-specific AS [61]. Also, hnRNP A1 and the SR proteins ASF/SF2 and SC35 have antagonistic functions in the splicing of the mammalian β -tropomyosin transcript [62].

In order to bind RNA and function in AS, SR proteins require the phosphorylation of their RS-domain [63]. A number of SR protein kinases, such as SR protein kinase 1

(SRPK1) and Clk/Sty kinase, are able to phosphorylate serine residues within the RS-domain [64, 65]. Modification of the RS-domain is also used as a means of regulating SR protein localization to the nucleus. Some SR proteins shuttle between the nucleus and cytoplasm, and reversible phosphorylation has been identified as one regulatory mechanism of this shuttling.

In addition to their role in regulating AS of pre-mRNA, SR proteins function in mRNA processing steps that occur after splicing. SR proteins, SRp20, 9G8, and ASF/SF2, have been shown to interact with the mRNA nuclear export receptor TAP/NFX1, and promote nuclear export of histone mRNAs [66, 67]. SR proteins have also been shown to influence translation. ASF/SF2 indirectly influences translation initiation by promoting the production of an isoform of the protein kinase, MNK2, which enhances cap-dependent translation [25]. ASF/SF2 is also involved in regulating translation directly through recruitment of components of the mammalian target of rapamycin (mTOR) signaling pathway [68]. SRp20 and 9G8 have also been implicated in increasing translation efficiency of viral and intronless mRNAs, respectively [69, 70]. In addition to translational regulation, SR proteins have been found to affect other RNA-processing steps. For example, the Nonsense-Mediated Decay (NMD) is a pathway that targets mRNAs containing premature termination codons for degradation. Increased expression of many SR proteins (ASF/SF2, SC35, SRp40, and SRp55) strongly enhances the NMD response [71]. SR proteins have also been implicated in transcriptional control. For example, the SR protein SC35 has been shown to associate with active chromatin, and affect transcriptional elongation by recruiting kinases to Pol II, independent of its role in AS [72]. These findings demonstrate that SR protein function is not restricted to nuclear

mRNA splicing, but play substantial roles in downstream mRNA metabolic functions in the cytoplasmic compartment.

1.9 Methods for the genome-wide detection of alternative splicing

Advances in technology have created the ability to explore and assay alternative splicing in a genome-wide manner. Various platforms for high-throughput detection of changes in AS have been developed, and contribute substantially to the identification of known splice variants in a given tissue, or that are changed under specific cellular conditions, like stress or stem cell differentiation [9, 73]. They have also been used to identify AS events regulated when a specific factor's expression is manipulated [74, 75]. There are two main technologies that are currently utilized to profile genome-wide splicing changes: splicing-sensitive microarrays and next-generation sequencing (also known as RNA-Seq).

A microarray consists of a glass slide with thousands to millions of oligonucleotides anchored to hybridize to transcripts that contain sequences of interest. Typically, RNA is isolated from a source (Tissue or specific cell type), and polyadenylated mRNA is enriched. Selected mRNA is subjected to reverse-transcription polymerase chain reaction (RT-PCR). The resulting cDNA is labeled with a fluorescent dye, hybridized to the microarray, and signal intensity is quantified at positions on the array where the sample is bound with the probe sequence. The intensity of the signal at each probe location on the microarray is proportional to the amount of mRNA that was in the sample. In order to detect changes in AS, microarray probes must be also designed to include splice junctions. Changes in probe intensities at the splice junctions for different experimental

conditions indicate AS changes. A limitation to using microarrays to detect changes in AS, is that profiling is restricted to annotated, or known, AS events. Also, due to the limited number of probe sets that can fit on a microarray, it is not possible to design probes to detect all possible AS outcomes. This limits the number of AS events that can be assayed, and prevents the discovery of novel, unannotated, AS events. Another drawback is the signal is measured indirectly, as a fluorescent signal, therefore differences in probe specificities, and other noise variables, can affect the readout [76, 77].

RNA-Seq is the newest high-throughput method to detect AS. RNA-Seq provides an approach to directly sequence millions of cDNAs generated from an RNA sample in a short amount of time, and at a small cost. There are many next-generation sequencing platforms used for RNA-Seq, and other high-throughput genomic applications. However, one of the most widely adopted platforms, and the platform used in the studies described in this dissertation, is Illumina sequencing. To prepare a sample for Illumina sequencing, mRNAs are selected by affinity purification using poly(T) oligos. Next, the mRNA is fragmented, and converted to cDNA using a random hexamer primer. After second strand synthesis, Illumina sequencing adapters are ligated onto the ends of the cDNA molecules, then the library is PCR amplified, and size selected at ~260bp. The resulting cDNA library is then sequenced starting from the ends of the molecule, and the number of base pairs sequenced depends on the number of sequencing cycles [78]. RNA-seq typically produces millions of short sequences known as “reads.” Currently, Illumina can sequence 35-100bp, but this technology is rapidly advancing and the total number of sequence reads, and the length of the read, is expanding exponentially.

In contrast to array-based methods, RNA-Seq does not rely on predesigned probes, and therefore can identify and quantify all the RNAs present in a sample, without the need to know the sequences beforehand. This allows for the identification of new AS events that were not known to exist [79]. This is also useful for profiling a transcriptome from an organism that is not well annotated [80]. A second advantage of RNA-Seq, is its low background signal, and consequently, greater dynamic range of expression detection [76, 78]. This allows for the detection of changes in gene expression and AS in low expressed transcripts.

Like many other high-throughput technologies, RNA-Seq faces several bioinformatic challenges. The data that is produced from an RNA-seq run is increasingly large, and several quality controls have to be implemented to reduce errors in base-calling, and removing low-quality reads [81]. One of the first steps of analyzing high-quality RNA-Seq data is to align the reads to a reference genome. Software has been developed to assist in aligning millions of short reads to the genome, rapidly. Bowtie is one alignment program that can align reads to the genome in a short amount of time [82]. One problem with Bowtie is that it allows mismatches, but not gaps in the alignment. While this is not a problem for most positions on the genome, reads that fall over splice junctions will not align using Bowtie. One solution is to create a database of known or predicted splice junctions, and then align the reads that did not align to the genome, to the database of splice junctions. As sequencing reads get longer, it will become more likely to have reads fall within the splice junctions, therefore a future challenge is to develop computational methods to simply align reads falling in splice junctions. Next, aligned reads are assigned to specific transcript isoforms, and then the number of reads within each isoform is

quantified. This allows for the quantification of changes in alternative splicing, or assaying changes in entire transcript isoforms [83-85].

RNA-Seq has clear advantages over previously developed high-throughput platforms, such as splicing-sensitive microarrays. It not only provides a more sensitive, unbiased, approach to analyzing AS, but also can be used to monitor other transcriptional changes. In addition to AS, information about transcript expression, stability, 5' and 3' end formation, can all be obtained from a single RNA-Seq experiment [78, 86].

1.10 Identification of the genome-wide RNA binding distribution of splicing factors

Multiple factors are important for the regulation of AS. RNA attributes, such as RNA structure and binding of noncoding RNAs, have been demonstrated to influence AS outcomes [87, 88]. However, protein-RNA interactions are considered the primary method of AS regulation, and to understand how splicing is regulated, it is important to know what RNA elements splicing regulators interact with to function in AS.

A number of computational approaches have been used to identify sequence motifs that regulate exon inclusion. The RESCUE-ESE (relative enhancer and silencer classification by unanimous enrichment) method analyzed motifs that occur within exons with a difference in splice site strengths, and identified 238 putative ESEs that occurred more frequently in exons with weak splice sites [89]. Another study compared exons with pseudoexons, and identified 2000 RNA octamers that were found more frequently in exons that are spliced [90]. More recently, bioinformatic studies have been performed on genome-wide data sets that identify RNA motifs that are enriched around alternative

exons that are regulated tissue specifically, or are changed in different developmental stages [83, 91]. Many of the putative ESE sequences were shown to enhance exon inclusion *in vivo*.

In conjunction with identifying regulatory sequences in the RNA, several approaches have been taken to identify the RNA motifs that are recognized by specific splicing regulators. One approach, SELEX (selected evolution of ligands through exponential enrichment), selects for high-affinity binding sites by assaying protein binding to a randomized pool of RNA sequences [92]. This has resulted in the identification of binding sites for many splicing regulators, including some members of the SR protein family [93-95]. Since these motifs are derived *in vitro*, an alternative approach, termed functional SELEX, was developed and allows for selection of binding sites from random sequences inserted into a mini-gene reporter construct. Functional SELEX allows for selection of sequences that regulate splicing, *in vivo*, rather than just RNA binding alone [96]. The enriched binding sites for splicing regulators derived from SELEX and functional SELEX were analyzed for their presence near alternative exons, and were used to predict exons regulated by several RNA-binding proteins.

The biochemical methods used to identify protein-binding sites were useful for predicting regulated exons, but these approaches did not assay RNA binding by native proteins in an intact cellular context. RNA immunoprecipitation (RIP) combined with RT-PCR or microarray technology was the first step to identify native protein-RNA interactions [97]. However, this approach is limited by low resolution of binding sites, and common identification of indirect interactions with the RNA. Additionally, reassociation of RNPs

after cell lysis complicated the ability to classify interactions as a true reflection of RNA binding, *in vivo* [98].

In order to increase binding site specificity and resolution, CLIP (cross-linking and immunoprecipitation) was developed. The CLIP procedure involves the *in vivo* UV cross-linking of RBPs to their target RNA molecules, this allows for purification of RNP complexes under stringent, denaturing, conditions [99]. When the resulting bound RNAs are subjected to next-generation sequencing (CLIP-Seq), genome-wide protein-RNA interactions can be identified [100]. This allows for the identification of, *in vivo*, specific and direct protein-RNA interactions that can reveal global principles of RBP function. This procedure was first used to characterize the RNA binding targets of the neuronal-specific splicing factor, Nova, but have since been used for many other RBPs, including the mammalian SR protein ASF/SF2 [101]. Although CLIP experiments produce data that is highly specific, cDNA libraries produced from CLIP have limited complexity. This is due to the inefficiency of UV cross-linking, and loss of complexity from inefficient steps in the library preparation [102]. Several modifications to the CLIP protocol have aimed at increasing the complexity, and specificity, of libraries being generated. Including, the use of photoreactive ribonucleoside analogs (PAR-CLIP), and modifications to the standard CLIP library preparation protocol that not only increases library complexity, but also allows for mapping of protein-RNA interactions at single-nucleotide resolution (iCLIP) [103, 104]. It was observed that many cDNAs prematurely end at the cross-linked nucleotide [105]. In order to capture the truncated cDNAs, the cDNAs generated from iCLIP are circularized, and then sequenced from the cross-linked nucleotide position. The use of a more efficient cDNA circularization step increased the

library complexity, and allows for the mapping of protein-RNA interactions at the single-nucleotide level. This technique has allowed for unprecedented resolution, and insight, on binding positions of many splicing regulators [106, 107]. Although many targets of regulators have been identified using one or more of these strategies, most of the targets of many RBPs are unknown.

1.11 Integrating genome-wide datasets: RNA splicing maps

Many early studies of AS focused on the regulation of a small number of splicing events, and characterized the mechanism of regulation of only a few splicing regulators. This means that many general conclusions about splicing regulation are supported by only a few examples. For example, SR proteins were thought to only enhance exon inclusion, and hnRNP family members silenced this enhancement. This general splicing conclusion was supported by analysis of minigene constructs of a few AS exons, using only a few members from each splicing factor family. However, over time, as SR and hnRNP proteins were studied more deeply, exceptions to this rule were uncovered [74]. SR proteins were found to have the ability to repress splicing when bound to certain intronic regions, and hnRNPs could enhance exon inclusion if bound downstream of the 5'ss [108, 109]. These results indicated that the binding position of splicing regulators on their targets determined how they functioned in regulation of AS. Therefore, to uncover the general principles of splicing regulation, multiple, genome-wide data sets need to be integrated.

Combining identification of genome-wide changes in AS that are regulated by an RBP, and the RBP's binding location on the target, yields splicing profiles that provide insight

into position-dependent regulation of AS (RNA splicing maps) [110]. The initial approach mapped the location of bioinformatically derived Nova binding sites around cassette exons that were found to be regulated by Nova using splicing-sensitive microarrays [111]. A comparison of binding location on the target transcript when Nova inhibited exon inclusion, and when it enhanced exon inclusion, revealed that Nova inhibited splicing when bound upstream of alternative exons, and activated inclusion when bound downstream. RNA splicing maps have since been generated using RNA-Seq and CLIP-Seq high-throughput datasets, and have detected the same primary positions of Nova binding near regulated exons [99, 100]. Similar RNA splicing maps have been generated for mammalian hnRNP family members, PTB, Fox 1/2, and the *Drosophila* Nova homologue Pasilla [83, 112-114]. Comparing the splicing maps of different RBPs, including comparing the maps of RBPs within the same family, can identify common principles of splicing regulation. Nova, Fox 1/2, and PTB can all silence exon inclusion by binding near the branch points or splice sites. RNA splicing maps for six members of the mammalian hnRNP family revealed that each hnRNP protein had a unique pattern of binding around regulated cassette exons to both activate or silence inclusion. The development of RNA splicing maps allows for the understanding of broad principles of splicing regulation, as well as, identifies proteins and targets that deviate from these basic models. In addition, building RNA splicing maps for multiple RBPs will identify relationships and interactions between multiple splicing regulators on the same pre-mRNA, and will reveal insights as to how combinatorial splicing control is achieved.

Further integrating splicing maps with transcription, chromatin dynamics, RNA structure, and post-translational modification information will allow for understanding of how an

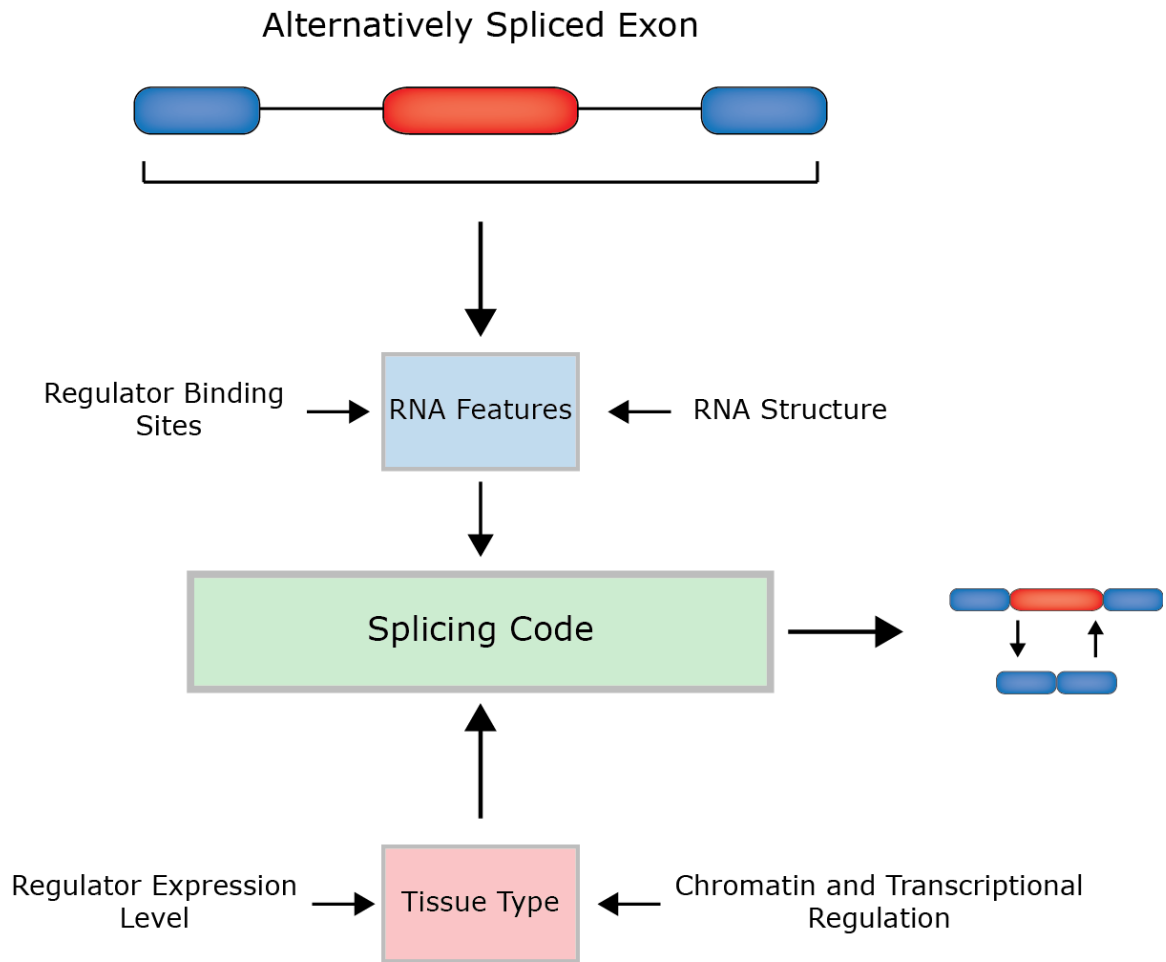


Figure 1.8: Uncovering the splicing code. The splicing code dictates the splicing pattern of a given pre-mRNA under a specific condition. RNA features, such as regulator binding sites and structure, and information about the location of splicing (tissue, stage, or disease state) are used to predict how splicing will be regulated.

exon will be spliced in a certain tissue, developmental state, or specific condition, what is termed the splicing code (Figure 1.8). The splicing code will describe how RNA features work together to regulate splicing in a specific context, and provide further insights into the mechanism of misregulation in disease.

1.12 Discussion

The precise control of AS by RBPs is essential for proper gene expression, and defects in regulation can result in severe genetic diseases and cancer. Our mechanistic understanding of AS has come from analyzing how *trans*-acting protein factors interact with *cis*-elements in target RNAs to regulate AS. More recently, it has been demonstrated that most AS events are regulated by multiple protein factors in a combinatorial, tissue specific, fashion. Therefore, understanding how multiple RBPs regulate the same AS event will provide clues as to how combinatorial regulation is achieved.

ASF/SF2 was the first SR protein identified to have activity in constitutive splicing, and the ability to modulate AS [42]. Additional RS-domain containing proteins, with an N-terminal RRM domain organization were identified, and termed SR proteins. Further studies demonstrated that other SR proteins could also function in splicing [115]. Most SR protein family members have been implicated in AS regulation, but many of the studies have been performed using *in vitro* systems, or modeled on single genes. There has never been a systematic genome-wide analysis of AS for a majority of the SR protein family members. Recent advances in the ability to rapidly characterize AS have allowed for a more comprehensive analysis of AS. It is now possible to use RNA-Seq to identify genome-wide AS events, and immunoprecipitation approaches to identify the precise binding location of RNA binding proteins across the transcriptome.

In summary, work from many different systems has provided insight into the molecular mechanisms of SR protein regulation of AS. It has also been shown that SR proteins have specific functions during embryonic development and tissue maintenance, and that

mutations and aberrant expression of certain SR proteins play significant roles in the development and progression of many human diseases. In the coming chapters, we utilize RNA-seq following RNAi (RNA interference) of the eight identified SR proteins in *Drosophila melanogaster*—SC35, SF2, SRp54, XL6, Rbp1, B52, Rsf1, Rbp1-like—to identify AS events regulated by individual SR proteins, and reveal that a majority of SR regulated AS events are regulated by multiple SR proteins in a combinatorial manner. Additionally, we find that SR proteins can regulate other RNA processing events such as transcript abundance, alternative promoter selection, and alternative polyadenylation. Next, we show that AS regulation requires an SR protein RS-domain, and the RRM-domain confers SR protein specificity. Using iCLIP-Seq, we identify direct RNA binding sites across the transcriptome for each SR protein, including position-dependent binding locations that contribute to the regulation of AS events. Finally, we integrate the two genome-wide data sets to produce RNA splicing maps for the SR protein family. By comparing the RNA bound by individual SR proteins, and the pre-mRNAs controlled by them, we will begin to define precise global regulatory networks for this family of splicing factors. From this analysis, and analyses of associated proteins, we will begin to understand the exact molecular mechanisms used by SR proteins to regulate AS. These findings will present new insights into gene expression regulation, and the contribution of AS to development and disease.

Chapter II: Identification and Quantification of SR Protein Regulated RNA Processing Events Using RNA-Seq

2.1 Introduction

Most of our understanding of how SR proteins regulate AS come from biochemical studies focused on single genes, and *in vitro* assays. These studies have provided a general understanding of the mechanisms of splicing regulation by the SR proteins, but only using a few SR protein family members, and a small number of alternatively spliced genes. Early studies of SR protein splicing regulation indicated that SR proteins enhance exon inclusion by binding to enhancing elements in the pre-mRNA, and recruiting core components of the splicing machinery, such as the U1 snRNP to the 5' splice site, and U2AF to the 3' splice site. In contrast, it was determined that hnRNP proteins interact with silencing elements to antagonize enhancers, such as the SR proteins. One of the best characterized example of this antagonistic regulation is the HIV-1 tat exon 3, where hnRNP silencing of splicing can be counteracted by an SR protein binding nearby [57]. However, exceptions to this model have been found. For example, SR proteins were found to repress splicing of the adenovirus L1 unit when bound to an upstream intron sequence, also hnRNP L enhanced exon inclusion when bound downstream of the 5'ss [108, 109]. An additional study comparing genes regulated by two SR proteins, ASF/SF2 and B52, and four hnRNP family members found that both SR and hnRNP could activate and repress AS events. Also, little overlap was found in regulated AS targets between SR and hnRNP family members [74, 116]. These studies suggest that SR proteins can both activate and silence exon inclusion in certain contexts, and that antagonistic regulation between SR and hnRNP proteins may not be a major mode of regulation.

SR proteins were first grouped together based on their protein domain organization, and ability to influence splicing. Initially, the ability of multiple SR proteins to enhance splicing suggested that SR proteins function as redundant regulators. However, the observation that embryonic or cellular lethality when a single SR protein is deleted could not be rescued by expression of other SR protein family members, suggested that SR proteins were not functionally equivalent [117]. Additionally, evidence that individual SR proteins have differences in their ability to regulate alternative splicing was shown. One of the earliest examples was that the overexpression of ASF/SF2 and SC35 produced different patterns of altered splicing of the adenovirus E1a pre-mRNA, but had a similar affect on SV40 early pre-mRNA splicing patterns [118]. These observations suggest that individual SR proteins are functionally distinct, and can regulate AS in specific ways.

In addition to AS, there are many other stages of RNA metabolism that SR proteins have a role in regulating. SR proteins have been shown to be involved in the regulation of transcription, translation, mRNA export, and mRNA stability. Primary transcripts can be initiated at alternative promoters (APR), and there can also be differential termination or 3' end processing (APA). Many eukaryotic genes contain multiple promoters that determine a different transcription start site and first exon of a transcript. This can affect the coding region of the protein, or it can affect regulatory regions in the 5' untranslated region (UTR) [119]. In addition to alternative start of a transcript, there are also changes in where a transcript terminates, known as alternative polyadenylation (APA). Like APR, APA can change the protein coding region of a transcript, or vary the length and composition of a transcript's 3' UTR [120]. Changing where a transcript begins and terminates has been shown to affect how transcripts are exported, localized, and the

efficiency of translation (protein expression) [121-123]. SR proteins have been shown to associate with components of the transcriptional machinery, including the Pol II CTD, and many of the other machineries that are required for these processing events, and may play a role in regulating the lengths of transcripts [124].

To gain insights into how AS is regulated by the entire family of SR proteins, it is important to first identify as many physiological target AS events that are affected by each protein. Historically, it has been difficult to identify endogenous AS targets for splicing regulators, but recent advances in splicing-sensitive microarrays and RNA-Seq allow for the discovery of genome-wide regulated AS events. In addition to changes in AS, RNA-Seq also allows for monitoring changes in overall gene expression levels, and changes in the first exon starting location (APR), and alternative polyadenylation. Using RNA-Seq when SR protein expression is knocked-down, we report here that SR proteins regulate a distinct, but overlapping, set of AS events. SR proteins can act as activators and silencers of inclusion of many AS events. Additionally, several SR proteins can regulate the same event, and in some cases, multiple SR proteins can regulate an AS event in opposite ways. We further show that this co-regulation is not redundant, but combinatorial. We also show that SR proteins can regulate the overall gene expression of transcripts, and can regulate APR and APA, independent of their role in canonical AS regulation.

2.2 Materials and Methods

2.2.1 Drosophila S2 cell culture

D.mel2 cells (GIBCO) were maintained at 25°C in *Drosophila* schneider's media (GIBCO # 21720) plus 10% fetal bovine serum (PAA #A15-701).

2.2.2 SR protein RNAi

In vitro transcription, RNA purification, and preparation of double-stranded RNA (dsRNA) were done as described [74, 116]. The dsRNA is generated from a PCR template carrying T7 RNA polymerase promoters on each end. Primers are designed carrying the T7 RNA polymerase promoter sequence and ~300-400nt of target sequence. For each SR protein two non-overlapping target sequences were chosen to produce two targeting dsRNAs for each protein using the primers below:

Primers used to generate dsRNA (5' to 3')

SC35-S-1

5-TAATACGACTCACTATAGGGAGGAATGAGCAACGGTGGTGGTGC-3

SC35-AS-1

5-TAATACGACTCACTATAGGGAGGAGCGGGCCATCTGTACGCGGA-3

SC35-S-2

5-TAATACGACTCACTATAGGGAGGACGCCCCTCTTCGCCCCTCG-3

SC35-AS-2

5-TAATACGACTCACTATAGGGAGGACTAGGAGCGACTGCGACTAC-3

SF2-S-1

5-TAATACGACTCACTATAGGGAGGAATGGGATCACGCAACGAGTG-3

SF2-AS-1

5-TAATACGACTCACTATAGGGAGGACCCACCGTCGCGGCTTCGGT-3

SF2-S-2

5-TAATACGACTCACTATAGGGAGGACCGCCAGCCAAGCGCTCGCAGTAC-3

SF2-AS-2

5-TAATACGACTCACTATAGGGAGGAGCCACCACCTCCACCGCTGCCG-3

SRp54-S-1

5-TAATACGACTCACTATAGGGAGGAATGGCTGGCGGCAACACCCC-3
 SRp54-AS-1
 5-TAATACGACTCACTATAGGGAGGACTTGAGCATCTCCAGGGCCCGAT-3
 SRp54-S-2
 5-TAATACGACTCACTATAGGGAGGATCACACCGCAGCGAGAAGCA-3
 SRp54-AS-2
 5-TAATACGACTCACTATAGGGAGGACTAGGGCGAGTTGGAGATGT-3
 XL6-S-1
 5-TAATACGACTCACTATAGGGAGGATCCGGAACTCGATCCCGCAG-3
 XL6-AS-1
 5-TAATACGACTCACTATAGGGAGGACTAGTCCCTTGAAACGGATC-3
 XL6-S-2
 5-TAATACGACTCACTATAGGGAGGAGCCGGGCTTCGCCTTCGTGG-3
 XL6-AS-2
 5-TAATACGACTCACTATAGGGAGGAGAGTTGCTTCTGCGTCGCTG-3
 Rbp1-S-1
 5-TAATACGACTCACTATAGGGAGGAATGCCGCGATATAGGGAGTG-3
 Rbp1-AS-1
 5-TAATACGACTCACTATAGGGAGGAAGGGCACGCGTTGCGTCTTC-3
 Rbp1-S-2
 5-TAATACGACTCACTATAGGGAGGAACACGCTGCTGCGGCACCAG-3
 Rbp1-AS-2
 5-TAATACGACTCACTATAGGGAGGATTAGAGTTGCAAGTTGAAGGTTGC-3
 B52-S-1
 5-TAATACGACTCACTATAGGGAGGAATGGTGGGATCTCGAGTGTA-3
 B52-AS-1
 5-TAATACGACTCACTATAGGGAGGAGTTGTAACGACCGCCCCCGC-3
 B52-S-2
 5-TAATACGACTCACTATAGGGAGGACGCTCCCGCTCGCGATCCAACAAA-3
 B52-AS-2
 5-TAATACGACTCACTATAGGGAGGACTAATCGTCCATGCTCTCGT-3
 Rsf1-S-1
 5-TAATACGACTCACTATAGGGAGGAGACTTTGGCCGGCACAGCAT-3
 Rsf1-AS-1
 5-
 TAATACGACTCACTATAGGGAGGATTAGAATCGATGGTTTCCCACCGGCGAG
 C-3
 Rsf1-S-2
 5-
 TAATACGACTCACTATAGGGAGGAATGTCCAGCATGGGTGATCAGCGCGGGA
 CA-3
 Rsf1-AS-2
 5-TAATACGACTCACTATAGGGAGGAGCGTCGTCCGCCCTGTCCA-3
 Rbp1Like-S-1
 5-TAATACGACTCACTATAGGGAGGAATGCCACGCTACCGTGAATGG-3
 Rbp1Like-AS-1

5-TAATACGACTCACTATAGGGAGGAGATGCGGGTGCCACAGCAGCG-3
Rbp1Like-S-2
5-TAATACGACTCACTATAGGGAGGAGTCGAAATGTCATCAGGCCGT-3
Rbp1Like-AS-2
5-TAATACGACTCACTATAGGGAGGACTAATGGCGGTCCCTGGAGTC-3

After PCR template is generated, *in vitro* transcription was performed using 50ng of PCR template, 150 Units T7 RNA polymerase (Invitrogen #18033-019) according to manufacture protocol, and incubated overnight at 37°C. RNA was purified using Qiagen RNeasy midi kit (Qiagen #75144). Purified RNA was annealed by diluting RNA to 1mg/ml in 10mM Tris-Cl pH 7.5, 100mM NaCl and heating to 95°C for 5 minutes then letting cool to room temperature for 1 hour.

RNAi knockdowns were performed by seeding 2×10^6 cells per well in a 6-well plate and adding 10µg dsRNA. After 48hr of incubation, a second round of 10µg dsRNA was added, and incubated an additional 48hr. The cells were harvested and the RNA purified using the RNeasy kit following the manufacturer's protocol (Qiagen #74106)

For experiments were two SR proteins were knocked-down simultaneously, 10µg of targeting dsRNA for each SR protein was added.

2.2.3 Generation of recombinant SR proteins for rabbit immunization

SR protein RRM-domains were amplified from cDNAs by PCR and cloned into a plasmid that contained an N-terminal pseudomonas endotoxin and a C-terminal 6x Histidine tag (pVCH6) using NheI/NotI restriction enzymes. One exception to this procedure was Rbp1-Like. Rbp1-Like was cloned using gateway technology (Invitrogen). First, Rbp1-Like RRM domain was PCR amplified, then it was cloned into an entry vector pDONR221 (Invitrogen) by BP reaction following standard protocols. Next, an

LR reaction was performed using pDONR221-Rbp1LikeRRM and pVCH6 that was converted to a destination vector. This results in the Rbp1-Like RRM domain inclusion in the pVCH6 plasmid.

Primers used for PCR amplification of SR protein RRM domain

SC35-RRM-S
5-GAATTCGCGGCCGCAATGAGCAACGGTGGTGGTGCC-3
SC35-RRM-AS
5-GTGATGGCTAGCGTAGCGGGCCATCTGTACGCG-3
SF2-RRM-S
5-GAATTCGCGGCCGCAATGGGATCACGCAACGAGTGC-3
SF2-RRM-AS
5-GTGATGGCTAGCTACGCGAATGTAGGCAACCTC-3
SRp54-RRM-S
5-GAATTCGCGGCCGCAATGGCTGGCGGCAACACCCCG-3
SRp54-RRM-AS
5-GTGATGGCTAGCGACGCTTAGATGGCCACCCTT-3
XL6-RRM-S
5-GAATTCGCGGCCGCAATGTCGCGCCATCCGAGCGAT-3
XL6-RRM-AS
5-GTGATGGCTAGCCACACGGGCTCGGCGCCCGCA-3
Rbp1-RRM-S
5-GAATTCGCGGCCGCAATGCCGCGATATAGGGAGTGG-3
Rbp1-RRM-AS
5-GTGATGGCTAGCTACGCGAATCCTGGTGCCGCA-3
B52-RRM-S
5-GAATTCGCGGCCGCAATGGTGGGATCTCGAGTGTAT-3
B52-RRM-AS
5-GTGATGGCTAGCGACCAAGTGGATCCGCCTGCC-3
Rsf1-RRM-S
5-GAATTCGCGGCCGCAATGTCCAGCATGGGTGATCAG-3
Rsf1-RRM-AS
5-GTGATGGCTAGCGCCAAAGTCGCCGCGTCGTCC-3
Rbp1Like-RRM-S
5-GCCAATGAATTCATGCCACGCTACCGTGAATGGGATTAGC-3
Rbp1Like-RRM-AS
5-CGTGTCGAGCTCGCGACCCTCTCGTGAACGGCCTGAT-3

Protein expression of the different recombinant proteins in Rosetta BL21(DE3)pLysS cells were done following standard procedures (Novagen). Following Lysis of the pelleted bacteria in denaturing lysis buffer (50mM Tris-Cl pH 7.5, 1M NaCl, 20 mM imidazole-HCL pH 7.5, 0.1% Triton-X, 6M Urea) proteins were purified using a 20 CV gradient of imidazole on a 1 mL HiTrap Chelating HP column (GE #17-0408-01) using an AKTA purifier 10 (GE Healthcare). The purified recombinant proteins were injected into two rabbits, per SR protein, according to standard protocols (Cocalico Biologicals, Inc, Reamstown, PA). After immunization, serum was extracted and SR protein specific antibodies were affinity purified.

2.2.4 Generation of recombinant protein and antibody affinity purification

Full-length SR proteins were amplified by PCR from cDNAs and cloned into the pET-30a(+) vector (Novagen) that contains a C-terminal 6x Histidine tag. Restriction enzymes NotI/NcoI for B52 and KpnI/NotI for the others were used for cloning. Again, Rbp1-Like was cloned using gateway technology (Invitrogen). First, Rbp1-Like was PCR amplified, then it was cloned into an entry vector pDONR221 (Invitrogen) by BP reaction following standard protocols. Next, an LR reaction was performed using pDONR221-Rbp1Like and pET-DEST42, which is pET-30a (+) converted to a Gateway destination vector. This results in the Rbp1-Like inclusion in the pET-DEST42 plasmid.

Primers used for PCR amplification of full-length SR proteins

SC35-FL-S

5- CAGATCGGTACCATGAGCAACGGTGGTGGTGCC-3

SC35-FL-AS

5- ACTCGATGCGGCCGCCTAGGAGCGACTGCGACTACG-3

SF2-FL-S

5- CAGATCGGTACCATGGGATCACGCAACGAGTGC-3
 SF2-FL-AS
 5- ACTCGATGCGGCCGCTTAATAGTTAGAACGTGAGCG-3
 SRp54-FL-S
 5- CAGATCGGTACCATGGCTGGCGGCAACACCCCG-3
 SRp54-FL-AS
 5- ACTCGATGCGGCCGCCTAGGGCGAGTTGGAGATGTC-3
 XL6-FL-S
 5- CAGATCGGTACCATGTGCGGCCATCCGAGCGAT-3
 XL6-FL-AS
 5- ACTCGATGCGGCCGCCTAGTCCCTTGAAACGGATCG-3
 Rbp1-FL-S
 5- CAGATCGGTACCATGCCGCGATATAGGGAGTGG-3
 Rbp1-FL-AS
 5- ACTCGATGCGGCCGCTTAGAGTTGCAAGTTGAAGGT-3
 B52-FL-S
 5- GTCCAGCCATGGATGGTGGGATCTCGAGTGTAT-3
 B52-FL-AS
 5- ACTCGATGCGGCCGCCTAATCGTCCATGCTCTCGTT-3
 Rsf1-FL-S
 5- CAGATCGGTACCATGTCCAGCATGGGTGATCAG-3
 Rsf1-FL-AS
 5- ACTCGATGCGGCCGCTTAGAATCGATGGTTTCCCAC-3
 Rbp1Like-FL-S
 5- GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGCCACGCTACCGTGAA-3
 Rbp1Like-FL-AS
 5-
 GGGGACCACTTTGTACAAGAAAGCTGGGTAATGGCGGTCCCTGGAGTCGGA-3

Protein expression of the different recombinant proteins in Rosetta BL21(DE3)pLysS cells were done following standard procedures (Novagen). Following Lysis of the pelleted bacteria in native lysis buffer (50mM Tris-Cl pH 7.5, 1M NaCl, 20 mM imidazole-HCL pH 7.5, 0.1% Triton-X) proteins were purified using a 20 CV gradient of imidazole on a 1 mL HiTrap Chelating HP column (GE #17-0408-01) using an AKTA purifier 10 (GE Healthcare).

~5mg of purified SR proteins were coupled to CNBr-agarose beads (GE Healthcare #17-0430-1) by first dialyzing the recombinant proteins 3x 1hr into coupling buffer (50mM

sodium borate pH 8.5 and 500mM NaCl). 0.5g of CNBr-agarose beads were activated, per sample, by rinsing with cold 100mM HCl for 20 minutes. Next, the beads were washed with 200mL of cold 100mM HCl, and then pelleted by centrifugation for 30 seconds at 750g. The protein sample was brought to 10mL (or more) with coupling buffer, and then added to the activated beads for 4 hours at room temperature. The remaining reactive groups were blocked by the addition of 1 ml of 1M Ethanolamine pH 8.0, and incubated an additional 1-hour. Beads were transferred to columns and washed 3x with buffer A (50mM Sodium Acetate pH 4.0, 500mM NaCl), 3x with buffer B (50mM Sodium Borate pH 8.0, 500mM NaCl), and 2x with PBS. The CNBr-agarose beads are now bound with recombinant protein.

The CNBr activated columns were used to affinity purify SR specific antibodies. First, the protein-bound beads were mixed with ~50mL of serum from SR protein immunized rabbits, and incubated overnight at 4°C. The serum/bead mixture is added back to the column and washed 1x with 10ml wash A (50mM Tris-Cl pH 8.0, 120mM NaCl, 0.5% NP-40), 1x with 10ml wash B (50mM Tris-Cl pH 8.0, 1M LiCl, 0.5% NP-40), again with 10ml wash A, then with 10ml PBS. The antibodies are eluted from the column into a fraction collector with 5ml elution buffer (50mM Glycine-Cl pH 2.5, 150 NaCl). Fractions were ~500µl and were neutralized with 7.5µl of 2M Tris-base. Fractions were analyzed by Bradford assay for protein presences and the most concentrated fractions were pooled as affinity purified SR protein antibodies.

2.2.5 Verifying antibody specificity and SR protein RNAi knockdown

After RNAi knockdown of individual SR proteins in S2 cells 10% of the cells were recovered added to NuPAGE LDS protein sample buffer (Invitrogen #NP0007). Samples were loaded on a NuPAGE 10% Bis-Tris midi gel (Invitrogen #WG1203BOX) and ran according to manufacture's protocol. The proteins were then transferred to a nitrocellulose membrane using the iBLOT transfer system (Invitrogen #IB301001). The membrane was blocked with 5% milk for 1hr, and then blotted with each SR protein affinity purified antibody (described above) at a dilution of 1:5000 for 1hr. The membrane was washed 3X 15 minutes with TBST. Next, rabbit secondary antibody conjugated with HRP (Invitrogen #G21234) was added at a dilution of 1:5000 for 1hr. The membrane was washed again 3x 15 minutes with TBST, then ECL western blotting detection agent (GE Life Sciences #RPN2132) was added and visualized with film.

2.2.6 Illumina next-generation sequence sample library preparation and sequencing

All sequencing libraries were prepared in collaboration with the Stowers Institute for Medical Research Molecular Biology Core facility using Illumina RNA-Seq sample prep kits (Illumina) according to manufacture's protocol. Briefly, polyadenylated RNA was purified from 2µg of total RNA using oligo-dT magnetic beads. The resulting RNA was fragmented, followed by first and second strand cDNA synthesis primed with random hexamers. The cDNA fragment ends were blunted (repaired) and Illumina sequencing adapter oligonucleotides were ligated to the ends. cDNA fragments are enriched by PCR amplification and size-selected from an agarose gel for a size of ~260bp. The library was validated and quantified on an Agilent Technologies 2100 Bioanalyzer using a DNA chip

(Agilent DNA-1000 kit). The samples were loaded onto a flow-cell for cluster generation, and sequenced using an Illumina Genome Analyzer II or Illumina HiSeq 2000 using single-read protocols (Illumina).

2.2.7 Alternative splicing event analysis (ASEA)

The ASEA pipeline was created in collaboration with Malcolm Cook and Marco Blanchette. Todd Bradley and Malcolm Cook wrote the description of the analysis. Models of transcription from Flybase are analyzed to produce an alternative splicing event (ASE) catalog for the *Drosophila melanogaster* genome (v5.22). For each gene in the genome, a pairwise comparison of all the gene's transcript isoforms is made. To identify alternative transcript isoforms that vary at the 5' and 3' ends, an ASE is declared wherever the comparison identifies divergent segments of transcription that are flanked on at least one side (5' or 3') by convergent transcription. The genomic coordinates of the segments unique to each form uniquely identify the event, and are entered into the catalog. Furthermore, the AS type (e.g. cassette exon) that characterizes the pattern of splicing is assigned based on the structure of transcriptional convergence/divergence, and can be used to sort ASEs by common types. This analysis produces a catalog of all the annotated ASEs in *Drosophila melanogaster*.

Bowtie is used to perform the alignment of sequence reads, from SR-depleted and control samples, to the *Drosophila melanogaster* (v5.22) genome, and a database of splice-junctions (36nt on each side of the splice-junction). Bowtie options allow for up to two-nucleotide mismatches, and only reads that produce a single alignment are kept. Reads that had more than two mismatches and/or aligned multiple times were discarded.

Next, for each event in the ASE catalog, the numbers of reads are tabulated that supports the inclusive form, exclusive form, both forms (constitutive), and reads that do not support either form. Two Fisher Tests are used to compare the tabulated read counts occurring in the SR-depleted and control results for each ASE. One is used to test significant changes in the inclusive counts, and the other tests for significant changes in the exclusive counts. This allows for identification of statistically significantly changed ASEs. To determine the magnitude and direction of splicing change PSI (percent spliced in) values are calculated. PSI is the ratio of counts of reads aligned to the inclusive region to counts of those aligned to either the inclusive or exclusive region. A PSI-switch score is computed as the difference between experimental and control PSI scores. A positive PSI-switch score indicates a greater presence of the inclusive isoform in the experimental sample (inclusion), whereas a negative score indicates greater presence of the exclusive isoform (skipping). Simple ASEs (cassette exon, alt. acceptor, alt. donor, retained-intron, and mutually exclusive exon) that were changed significantly ($p \leq 0.05$), and changed by at least 10% ($\text{PSI-switch} \geq 0.1$) were included for down-stream analysis. In addition, the same filter was used on ASEs that differed in their 5' and 3' ends.

2.2.8 RT-PCR validation of affected targets

cDNA was generated by reverse transcribing $\sim 5\mu\text{g}$ of the total RNA extracted from the control and SR-depleted cells using a mixture of oligo (dT) and random hexamer (N_6) primers, and Superscript III reverse transcriptase (Invitrogen) according to the manufacturer's protocol. PCR amplification of predicted targets was performed for 25-30 cycles with primers falling in exons flanking the alternative spliced region.

Primers used for PCR of ASEs:

CG34232_F
5-TTGCATCGCAGAAACAGAAG-3
CG34232_R
5-GTTCAGATGATGGGTTCGCT-3
CG12090_F
5-AGTCTTACCTCTGGGGAGCC-3
CG12090_R
5-AAGTAGTCCGTGGTGATGGG-3
CG31638_F
5-GGGAGTTAGCTGGTGCTGAG-3
CG31638_R
5-CTAGTCGCTGTTACCCTCCG-3
Sfmbt_F
5-CGAGCAGTCAACGACACAGT-3
Sfmbt_R
5-CACCAGCTCCAAGTGAGGTT-3
Vps35_F
5-TCGGCAATCGATATCTCCTC-3
Vps35_R
5-CTCGCTAGTCCAATGGCTTC-3
Nfat_F
5-CGCAGCAGCATTTAATCAGA-3
Nfat_R
5-GTTCAGCGGTGTAGTGGGTT-3
Su(Z)12_F
5-AGTATGGCCAGAAACGCAAT-3
Su(Z)12_R
5-GCCGCTTGTTAAGCACTGTA-3
Syb_F
5-TGATCATTCTGGGCGTGATA-3
Syb_R
5-TTTGGTCCTTTTCCATTTCG-3
Dre4_F
5-GCTCTGATGAGGAATCTGGC-3
Dre4_R
5-GTCCCGGCTGTTGTACTTGT-3
CG12065_F
5-CAAGGAGACATTCGTCCGAT-3
CG12065_R
5-GTCGACCTTTTGAATGTGC-3
DppIII_F
5-GCACTTTTGTCTGGCATTCT-3
DppIII_R
5-CTCGGTGAGATTCTCGAAGG-3

Pep_F
5-GTTCGGATCGGAGCGTAATA-3
Pep_R
5-CGGATTTCCGTTGACCTTTA-3

PCR products were analyzed on an Agilent Technologies 2100 Bioanalyzer, using a DNA chip (Agilent DNA-1000 kit) according to manufacture's instructions.

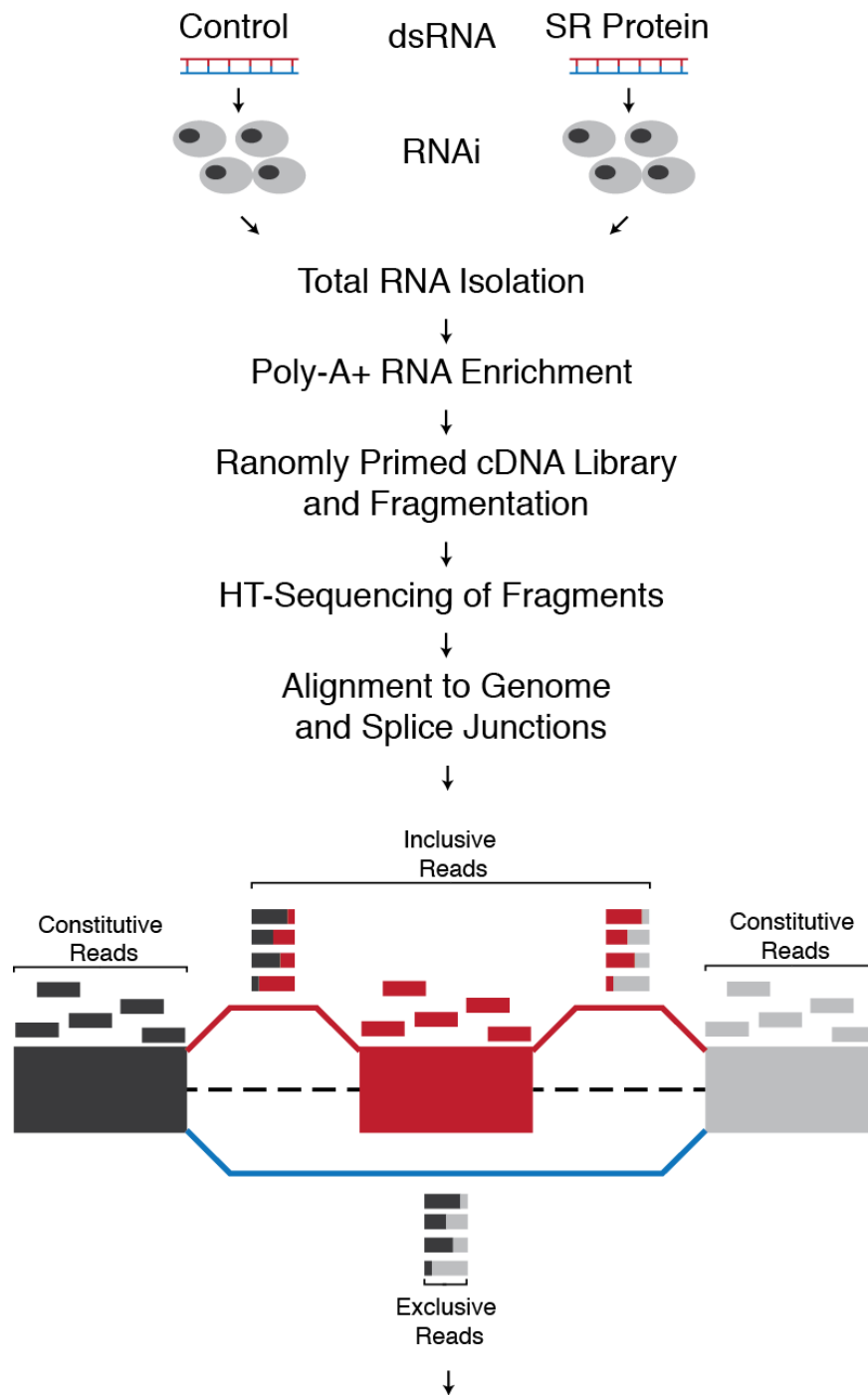
2.2.9 Identifying changes in overall transcript abundance

RNA-Seq reads from control and SR-depleted experiments were processed and aligned, simultaneously, to the genome and splice-junction database as described previously. For each condition, the mRNA RPKM (reads per kilobase of million mapped reads) was computed for each gene to measure relative expression. To detect differentially expressed genes in SR-depleted samples, the bioconductor edgeR package was used. EdgeR provides methods for the statistical analysis of count data from comparative experiments, using RNA-Seq data. Over-dispersed Poisson count models are used to distinguish true biological variation from technical variation. This allows for the identification of significant differentially expressed genes in the SR-depleted samples. We limited our analysis to genes that were in the top three quartiles of expression (top 75%) based on their RPKM values, and only considered significantly changed genes that varied ≥ 2 -fold.

2.3 Results

2.3.1 Identification of genome-wide SR-dependent alternative splicing events

To identify the SR protein network regulating AS, we reduced the protein levels of SR proteins SC35, SF2, SRp54, XL6, Rbp1, B52, Rsf1, and Rbp1-like using RNAi. Each protein was knocked-down, in duplicate, using two non-overlapping dsRNAs (Figure



$$\text{PSI}(\Psi) = \frac{\text{Inclusive Reads}}{\text{Inclusive Reads} + \text{Exclusive Reads}}$$

$$\text{PSI}(\Psi)_{\text{Switch}} = \Psi_{\text{Experimental}} - \Psi_{\text{Control}}$$

Figure 2.1: Experimental strategy for identification of SR protein regulated AS events. Total RNA was extracted from *Drosophila* S2 cells treated with two non-overlapping SR protein-targeting dsRNAs or control nontargeting dsRNAs. Next, RNA-Seq libraries were prepared, and subjected to Illumina next-generation sequencing. The resulting sequence reads were aligned, simultaneously, to the genome and a splice-junction database. Then, using a catalog of ASEs, the aligned tags were assigned to the inclusive or exclusive isoforms. A student's t-test was performed to determine ASEs that significantly changed in an SR protein knock-down compared to control, and a tabulation of reads within each isoform produced PSI and PSI-switch scores to quantify the splicing changes.

2.1). To assay protein levels of specific SR proteins, antibodies were produced against the RRM-domain of each individual SR protein in rabbits, and western-blot were performed on all samples (Figure 2.2). From this analysis, we were able to confirm that each antibody specifically recognized each individual SR protein with no cross-reactivity with other members of the family. Additionally, all SR proteins levels were reduced to less than 10% of their endogenous levels in *Drosophila melanogaster* S2 cells. Individual SR proteins were reduced specifically, with no appreciable changes in other SR protein family members (Figure 2.2). One exception was the previously identified Rbp1-like up-regulation when Rbp1 expression is reduced [125]. We next performed RNA-Seq on RNA isolated from control and SR RNAi-treated cells to detect altered splicing events in an unbiased fashion (Figure 2.1). We generated ~23-40 million sequence reads per sample that were aligned simultaneously to the *Drosophila melanogaster* genome (v5.22), and a splice junction database containing 36 bp on each side of the exon-exon junction, using Bowtie. This resulted in the unique alignment of 66.4% – 91.1% of the reads to the genome or splice junctions (Figure 2.3). Using a catalog of all annotated AS events in *Drosophila melanogaster* (genome version 5.22) aligned reads were assigned to

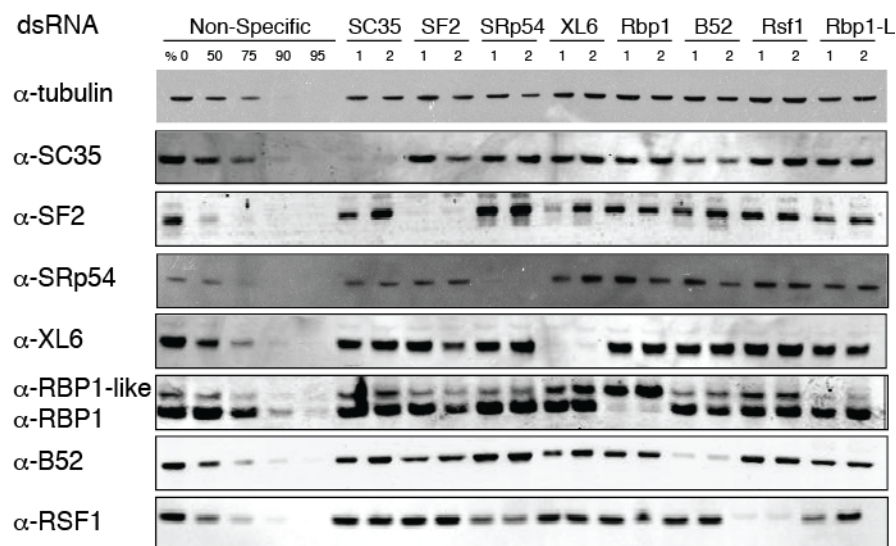


Figure 2.2: RNAi of the eight members of the *Drosophila* SR protein family in *Drosophila* S2 cells. Western-blot analysis demonstrating specific depletion of individual SR proteins in *Drosophila melanogaster* S2 cells, in duplicate, as compared with cells treated with a non-specific dsRNA. Tubulin was used as a loading control.

the inclusive, exclusive, and constitutive isoforms for each alternative splicing event (ASE) in the catalog, and tabulated (Figure 2.1). Reads that aligned to the ASE region, but did not support any of the known forms were grouped separately, and excluded from the further analysis. These reads may support previously unidentified transcript isoforms. This was done independently for reads for each SR protein knock-down and control samples. The ratio of the counts of reads that aligned to the inclusive isoform to counts of those aligned to either the inclusive or exclusive isoforms are computed for each ASE. This is referred to as the percent spliced in (PSI) score (Figure 2.1). The change in PSI scores between the experimental and control samples are calculated as the PSI-switch score (Δ PSI). This score indicates the direction (more inclusion, less inclusion), and the

Individual RNAi	Total Sequence Reads	Total Aligned Reads	Percent Aligned
Control-1	23,748,338	20,600,055	86.7%
Control-2	29,804,748	28,318,715	73.3%
SC35-1	25,350,320	20,854,701	82.3%
SC35-2	38,238,997	29,253,771	76.5%
SF2-1	40,332,728	34,788,279	86.3%
SF2-2	36,701,829	26,084,505	71.1%
SRp54-1	33,304,516	30,327,540	91.1%
SRp54-2	38,622,062	28,318,715	73.3%
XL6-1	28,968,252	26,249,598	90.6%
XL6-2	34,549,662	25,614,113	74.1%
Rbp1-1	38,226,070	32,568,940	85.2%
Rbp1-2	39,470,556	26,295,632	66.6%
B52-1	36,433,569	30,254,907	83.0%
B52-2	33,973,834	22,963,151	67.6%
Rsf1-1	40,516,480	33,387,729	82.4%
Rsf1-2	30,133,723	19,993,797	66.4%
Rbp1-like-1	37,068,460	31,170,472	84.1%
Rbp1-like-2	37,155,950	25,579,732	68.8%
Simultaneous RNAi			
Control-1	105,258,840	81,355,418	77.3%
Control-2	78,513,137	60,672,417	77.3%
SC35-B52-1	99,642,196	75,262,976	75.5%
SC35-B52-2	89,805,452	67,834,273	75.5%
XL6-B52-1	80,590,717	57,893,884	71.8%
XL6-B52-2	87,814,883	65,645,158	74.8%

Figure 2.3: RNA-Seq of RNA extracted from control and SR-depleted S2 cells. Total number of sequence reads from each sample, and the total number of reads that aligned to the *Drosophila melanogaster* genome and splice junction database.

magnitude of splicing change observed when an individual SR protein is reduced. To evaluate if an ASE is significantly changed upon reduction of an individual SR protein, a fisher's exact test is used to compare the tabulated read counts occurring in the control and SR protein knock-down sample. We used a stringent criterion of, $p \leq 0.05$ and a $\Delta\text{PSI} \geq 0.1$, this identifies statistically significant AS changes, which change in magnitude by at least 10%.

To simplify our analysis, we focused on the five simplest types of AS (cassette exons, competing donor or acceptor sites, mutually exclusive exons, and intron-retention). Using our criterion, we identified 561 altered simple AS events upon SR protein depletion, representing 405 genes. This represents more than 10% of simple AS events that can be reliably detected in S2-cells (5,472 events). Individually, SR proteins regulate different numbers of events (Figure 2.4A). B52 regulates the largest number of AS events (253) and Rbp1 the smallest (24). Previous studies of SR proteins focused on their regulation of cassette exons or competing 5' and 3' ss. To test if SR proteins are important for only one or a few types of simple AS, we subdivided the affected events into each type of simple AS. SR proteins were found to regulate all types of simple AS events, and affect different types of simple ASEs in similar proportions (Figure 2.4B).

SR proteins have previously been shown to be essential for proper tissue development, and sex determination in *Drosophila melanogaster* [50, 61, 126]. Gene ontology (GO)

A

	B52		SRp54		SC35		SF2		XL6		Rsf1		Rbp1		Rbp1-L	
	inc.	exc.	inc.	exc.	inc.	exc.	inc.	exc.	inc.	exc.	inc.	exc.	inc.	exc.	inc.	exc.
Cassette Exon	34	30	6	46	9	10	11	14	7	8	5	5	2	2	2	2
Alternative Acceptor	20	4	9	16	5	4	8	8	3	15	3	1	2	1	5	1
Alternative Donnor	31	9	23	11	15	6	17	11	6	18	14	5	4	2	6	2
Retained Intron	68	19	43	19	27	5	30	7	5	27	17	7	6	3	12	4
Mutually Exclusive Exon	25	13	8	16	3	6	1	6	0	5	0	2	0	2	0	1

B

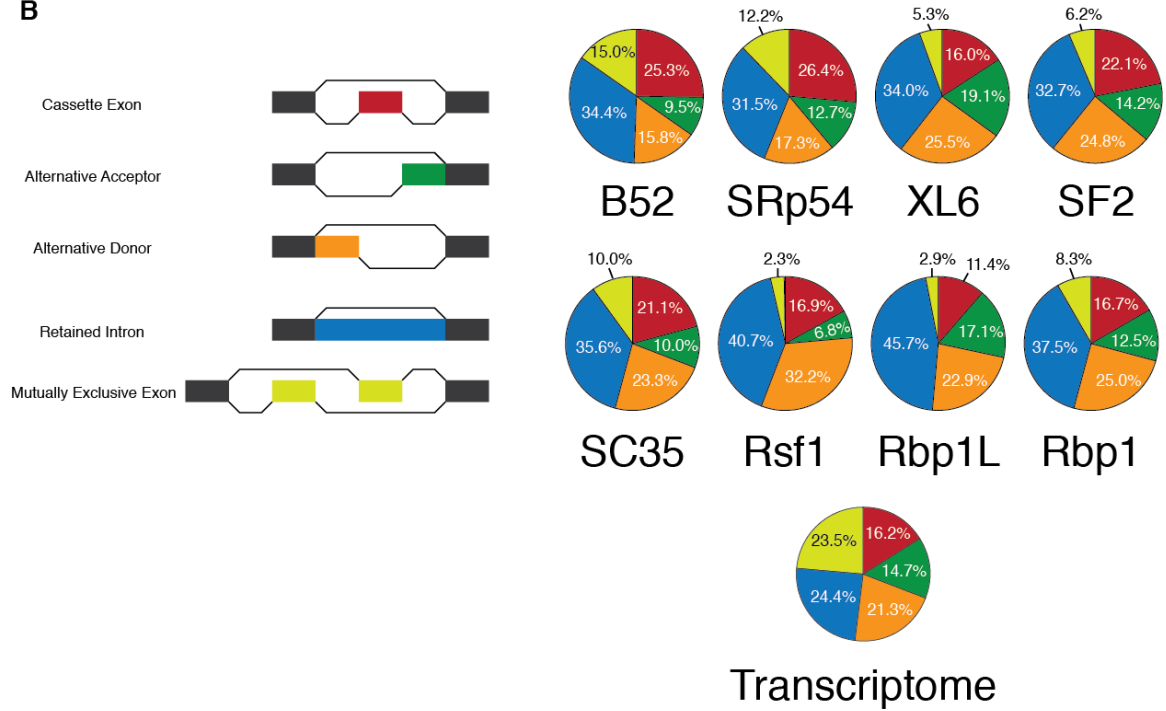


Figure 2.4: SR proteins regulate all types of simple AS events. (A) Total number of AS events regulated by individual SR proteins separated by causing more inclusion (inc.) or exclusion (exc.) of the alternative region. Number of events are further divided into the five simple AS types. **(B)** Pie charts of the distribution of simple AS types differentially regulated by individual SR proteins detected by RNA-Seq. Colors represent the different simple AS types that were detected, and the transcriptome is the distribution of simple AS events for the expressed transcriptome.

analysis of SR regulated targets using TopGO (using the top 40 nodes) revealed overrepresented terms corresponding to cell growth (cell growth, growth, and cell proliferation), cellular signaling (regulation of signaling, cell communication), tissue and organ generation (system development, tissue and organ development), and neuronal processes (generation of neurons, behavior, and neuron differentiation) (Figure 2.5). Thus, SR proteins regulate targets important in a broad range of physiological processes.

2.3.2 SR proteins can function as activators and repressors of splicing events

Previous studies have suggested that SR proteins bind to ESEs in the pre-mRNA, and function as general activators of exon definition. To investigate whether SR proteins generally activate or repress AS events, we determined the number of AS events that showed increased inclusion or exclusion when an individual SR protein was depleted. We found that all SR proteins can function as, both, alternative splicing event activators and alternative splicing event silencers (Figure 2.6A and B). Surprisingly, we found that SR proteins are important for the repression of more than half of their regulated events (58.3% for Rbp1 to 71.4% for Rbp1-like events), with the exception of SRp54 and XL6, which are important for the activation of 54.8% and 77.7% of their events, respectively (Figure 2.6A). In order to determine if SR proteins preferentially activate or repress a

Gene ontology analysis of genes with changes in alternative splicing affected by SR protein depletion in S2 cells

GO.ID	Term	Fisher.Test
GO:0050789	regulation of biological process	1.20E-13
GO:0050794	regulation of cellular process	5.80E-13
GO:0065007	biological regulation	7.70E-13
GO:0016043	cellular component organization	4.60E-12
GO:0071840	cellular component organization or bioge..	1.80E-11
GO:0048513	organ development	5.60E-11
GO:0040007	growth	6.70E-11
GO:0071842	cellular component organization at cellu...	3.30E-10
GO:0071841	cellular component organization or bioge..	1.00E-09
GO:0009987	cellular process	1.90E-09
GO:0009653	anatomical structure morphogenesis	6.40E-09
GO:0048699	generation of neurons	8.00E-09
GO:0048731	system development	1.10E-08
GO:0048468	cell development	1.50E-08
GO:0048856	anatomical structure development	2.60E-08
GO:0032501	multicellular organismal process	3.10E-08
GO:0032502	developmental process	3.60E-08
GO:0023052	signaling	4.20E-08
GO:0007275	multicellular organismal development	5.00E-08
GO:0007154	cell communication	6.10E-08
GO:0040011	locomotion	6.70E-08
GO:0002165	instar larval or pupal development	7.30E-08
GO:0048609	multicellular organismal reproductive pr...	7.80E-08
GO:0007610	behavior	8.00E-08
GO:0009791	post-embryonic development	8.00E-08
GO:0051128	regulation of cellular component organiz...	9.50E-08
GO:0023051	regulation of signaling	1.30E-07
GO:0007552	metamorphosis	1.40E-07
GO:0032504	multicellular organism reproduction	1.60E-07
GO:0030154	cell differentiation	1.80E-07
GO:0009886	post-embryonic morphogenesis	1.90E-07
GO:0030036	actin cytoskeleton organization	2.10E-07
GO:0048869	cellular developmental process	2.20E-07
GO:0009888	tissue development	2.30E-07
GO:0030182	neuron differentiation	2.50E-07
GO:0008283	cell proliferation	2.70E-07
GO:0030029	actin filament-based process	2.80E-07
GO:0007276	gamete generation	2.90E-07
GO:0048707	instar larval or pupal morphogenesis	3.10E-07
GO:0022607	cellular component assembly	3.60E-07

Figure 2.5: Gene ontology analysis of SR regulated genes. Significantly enriched GO-terms of genes regulated by SR proteins at the level of AS, using TopGO. Terms are listed in descending order of p.value as determined by a fisher-test. Only the top 40 nodes are displayed.

specific AS type, we separated the included or excluded events by AS type (Figure 2.7 A-E). SR proteins tend to activate more cassette exons than other types of AS, with the exception of B52, which represses 53.1% of cassette exon events (Figure 2.7A).

Additionally, SR proteins are important for the repression of most intron-retention splicing events, except for XL6, which activates 84.4% of its regulated intron-retention events (Figure 2.7D). Since SR proteins affect more intron-retention events than any other form of AS, these events contribute significantly to the overall repression of AS observed (Figure 2.4B). The observation of cassette exon activation, and repression of intron-retention events, are consistent with reports that SR proteins activate exon recognition [54]. However, there are many examples where SR proteins silence inclusion of AS regions, and to different extents.

Semi-quantitative RT-PCR of predicted targets demonstrates that multiple SR proteins can act as activators (*Su(Z)12* and *CG12090*) or repressors (*Pep* and *Syb*) of the same splicing event (Figure 2.8). Surprisingly, we also identified events where some SR proteins activate inclusion, and others act to repress inclusion. Such as the enhanced skipping of a cassette exon in the *Dre4* transcript when XL6 or Rsf1 are depleted, and the increased inclusion of the same exon when SC35 or B52 are depleted (Figure 2.8). This observation of antagonistic regulation suggests that all SR protein AS regulation is not completely redundant, but different SR proteins can regulate an alternatively spliced

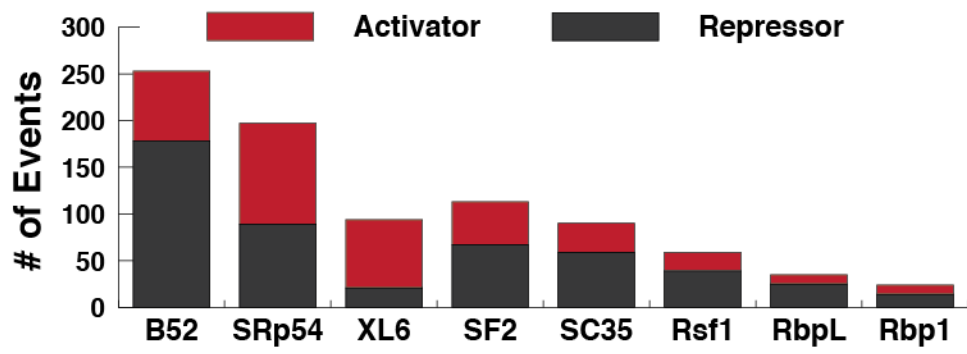
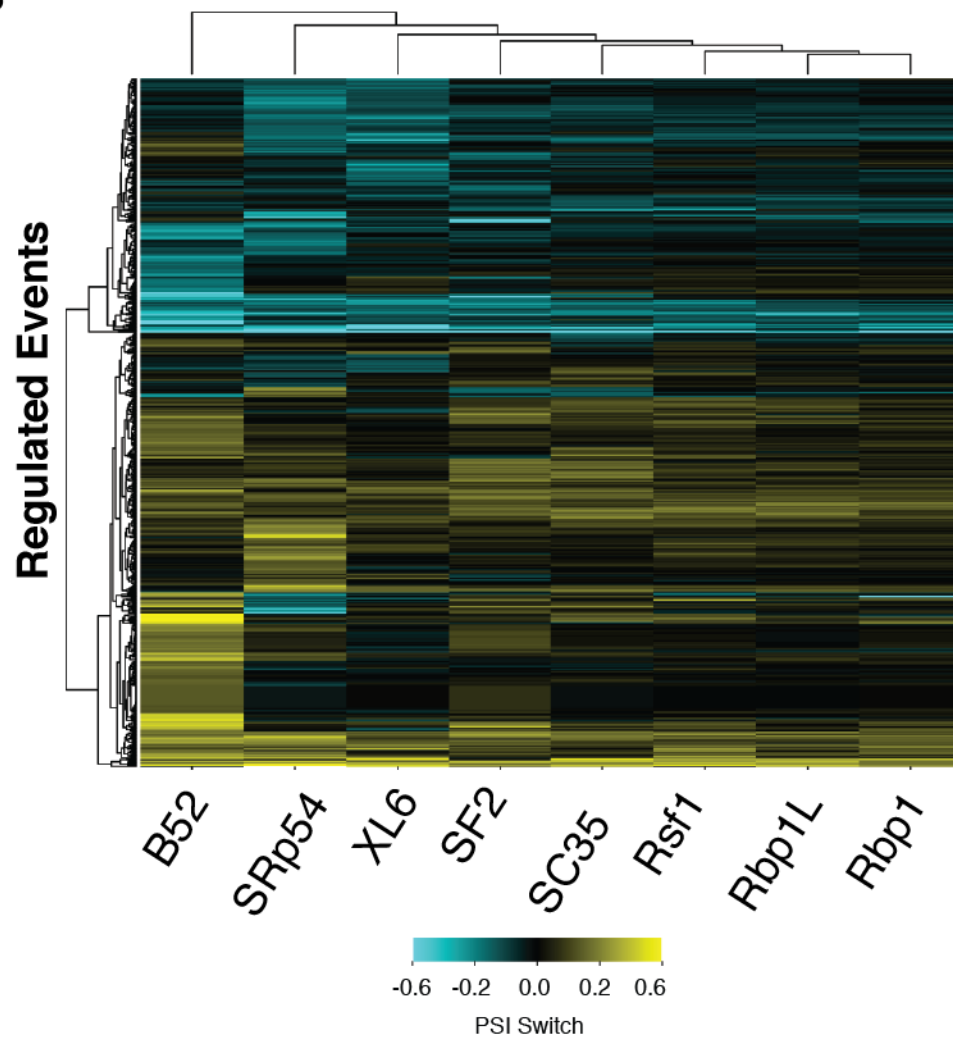
A**B**

Figure 2.6: Genome-wide analysis of SR-dependent AS events. (A) Number of simple AS events regulated by each SR protein. Bars are divided to display the proportion of events that are less included upon depletion (SR-activated, red), or more included upon depletion (SR-repressed, dark gray). **(B)** Heatmap shows hierarchical clustering of PSI-switch scores for SR-regulated AS events.

region differently. We conclude that all 8 SR proteins can act both as repressors and activators of AS events. Additionally, multiple SR proteins can regulate the same event, and in some cases multiple SR proteins can regulate an AS event antagonistically.

2.3.3 SR proteins collaborate to regulate a majority of AS events

Many alternative splicing events have been demonstrated to be regulated by multiple RNA splicing factors in a combinatorial manner [116]. For example, the SR proteins Rbp1 and SF2 cooperate with Tra and Tra-2 to regulate the *Drosophila* AS event in *Doublesex* [61]. Additionally, cooperative regulation, genome-wide, of AS by another well-characterized splicing regulator family, hnRNPs, has been demonstrated in *Drosophila* and mammals [74, 112]. Hierarchical clustering across SR protein regulated AS events demonstrates that SR protein proteins co-regulate many of the same targets (Figure 2.6B). We found that while individual SR proteins can regulate unique AS events, a majority of events are regulated by more than one SR protein (Figure 2.9A). One notable exception is B52, which shares only 36% of its targets with other SR protein family members. To determine if SR proteins are acting in concert or antagonistically, we noted the direction of change for each co-regulated event, and clustered them according to similarity (Figure 2.9B). Interestingly, a vast majority of co-regulated events are affected in the same direction (greater than 87.7% in every case), and very few

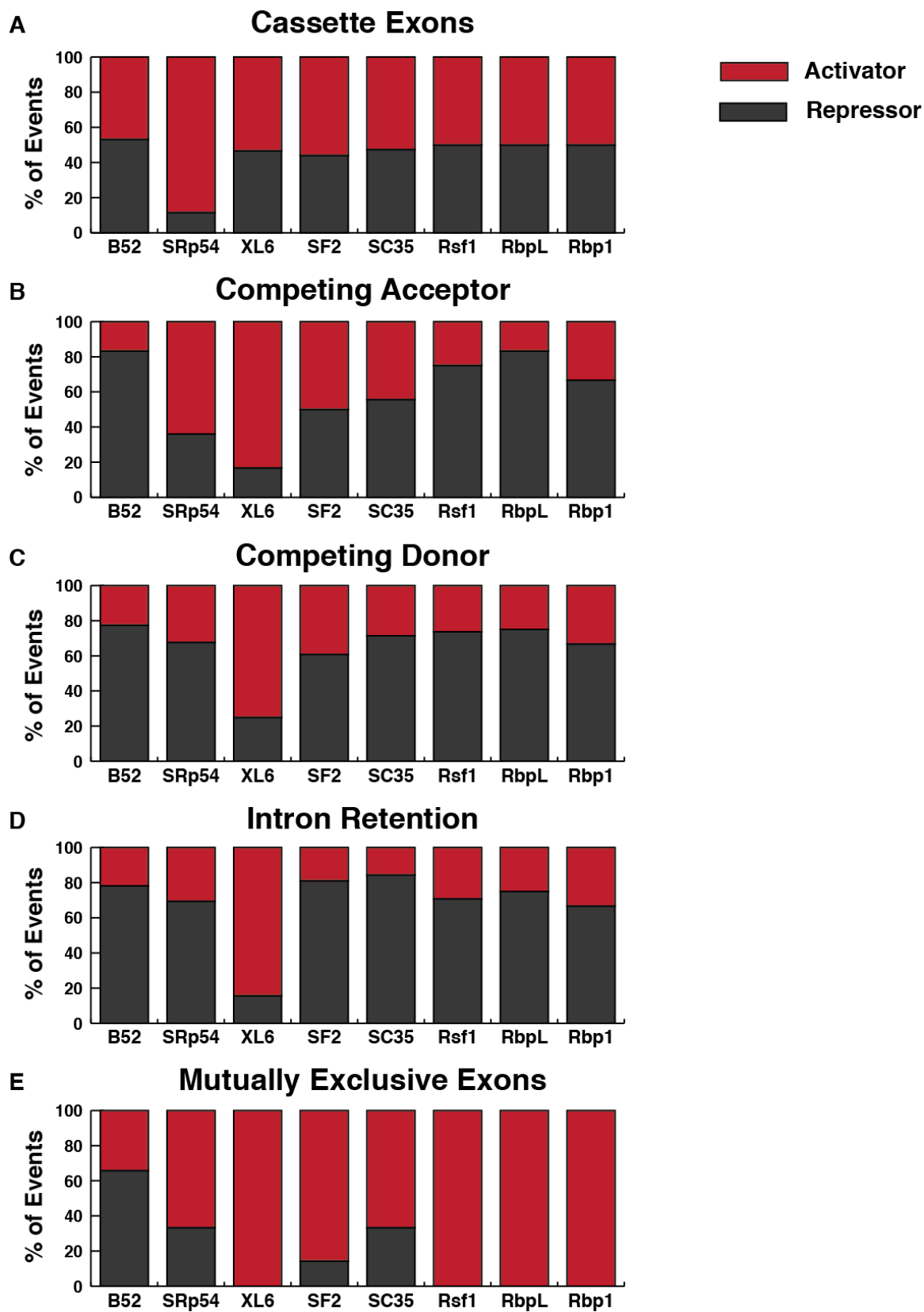


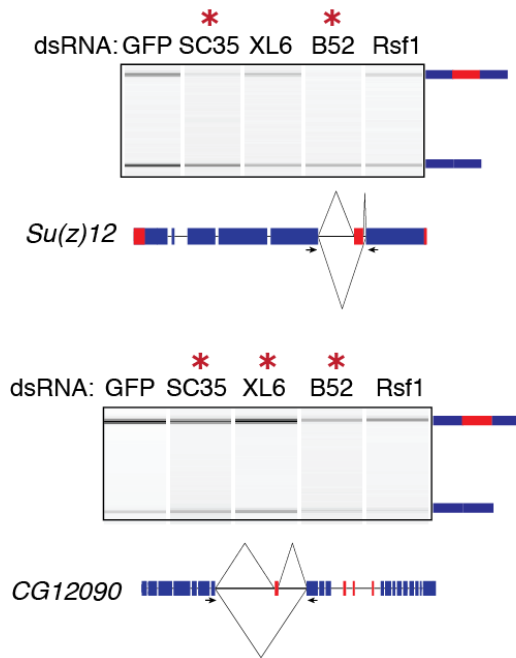
Figure 2.7: Proportion of events that are less included upon depletion (SR-activated, red), or more included upon depletion (SR-repressed, dark gray) by AS type (A-E).

events are regulated antagonistically. Also, while individual events can be regulated by specific sets of SR proteins, there are no distinct pairwise relationships between individual SR protein family members (Figure 2.9B).

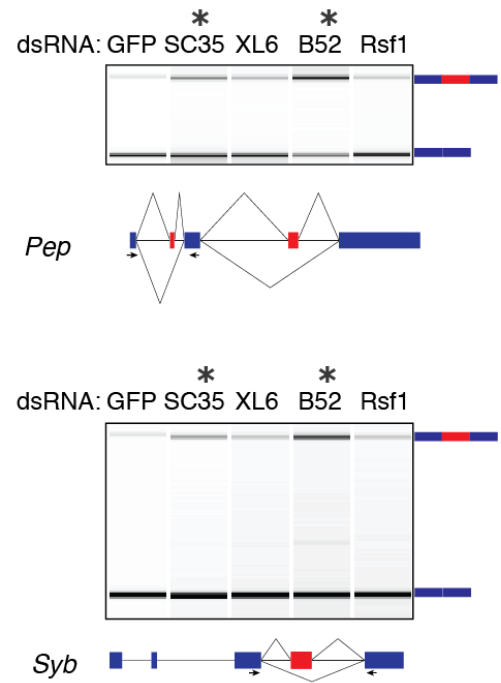
Rsf1 (repressor of splicing 1) has an RRM domain that shares 54%-58% sequence identity with other SR proteins, but has a C-terminal domain that is rich in glycine, arginine, and serine (GRS-domain). Previous studies have demonstrated that Rsf1 acts as an antagonist of other SR proteins for splice site selection through protein interactions within its GRS-domain [127]. Surprisingly, we found that Rsf1 co-regulates 81.4% of its events with other SR proteins, and 97.9% of those events are regulated in the same direction (Figure 2.9A and B). Therefore, Rsf1 acts in concert with the other SR protein family members in most of the identified co-regulated AS events, and is not a global repressor of AS through SR protein competition.

The extensive overlap of regulated AS transcripts raises the possibility that SR proteins may redundantly regulate specific AS events. Another possibility is that the SR proteins may regulate AS events using a combinatorial mode of regulation. To provide insights into the mechanism of SR protein co-regulation, we reduced the expression of two SR proteins simultaneously (SC35:B52 and XL6:B52) in S2 cells. We confirmed that both SR protein expression levels were reduced by western-blot (Figure 2.10A). We next isolated the RNA and prepared libraries for RNA-Seq (described above). After

Activator



Repressor



Antagonistic

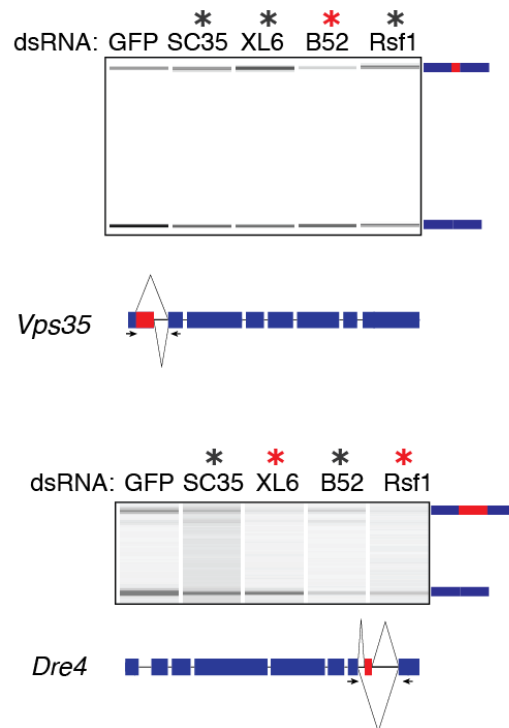
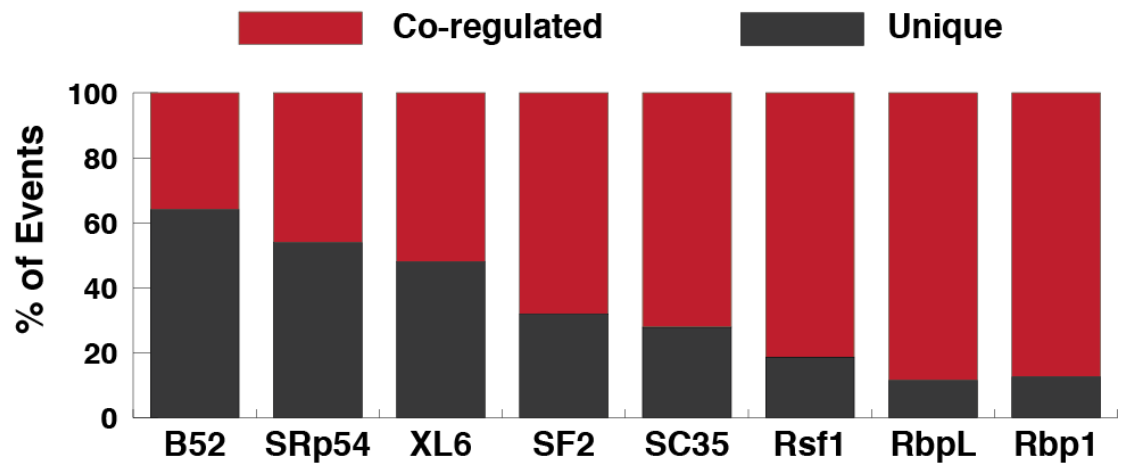


Figure 2.8: Differentially regulated events detected by RT-PCR. RT-PCR validation of SR-regulated AS events ($p \leq 0.05$; $\Delta\text{PSI} \geq 0.1$). Events separated into activator, repressor, or antagonistic. Red color stars indicate activator and gray stars indicate repressor.

sequencing, we received ~70-105 million reads that aligned to either the *Drosophila* genome (v5.22), or our splice junction database at ~72-77% (Figure 2.3). Next, we identified changes in AS between control and simultaneous SR RNAi samples, as described previously for individual RNAi samples. In addition to the regulated AS events identified when only an individual SR protein is reduced, 205 and 195 new AS event changes were detected in the SC35:B52 (395 regulated AS events) and XL6:B52 (347 regulated AS events) samples, respectively (Figure 2.10B). The identification of so many new events suggested that SR proteins are regulating AS in a combinatorial fashion, as well as the potential identification of redundantly regulated events. Also, we noticed that the overall magnitude of AS changes were greater for most of the events when two SR protein levels are reduced (Figure 2.10B). To examine this further, we compared the PSI-switch scores of the 23 co-regulated AS events between SC35 and B52, and the 26 co-regulated AS events between XL6 and B52 to determine if there was an increase in the magnitude of splicing change when both SR proteins were knocked-down (Figure 2.10C). In 14 of the 23 events co-regulated by SC35 and B52, there was an increase in the magnitude of splicing change (PSI switch score) when both SR proteins were knocked down simultaneously. The same was observed for 12 out of the 26 events regulated by XL6 and B52 (Figure 2.10C). This combinatorial regulation can be further demonstrated by semi-quantitative RT-PCR (Figure 2.11). The AS cassette exon in

A



B

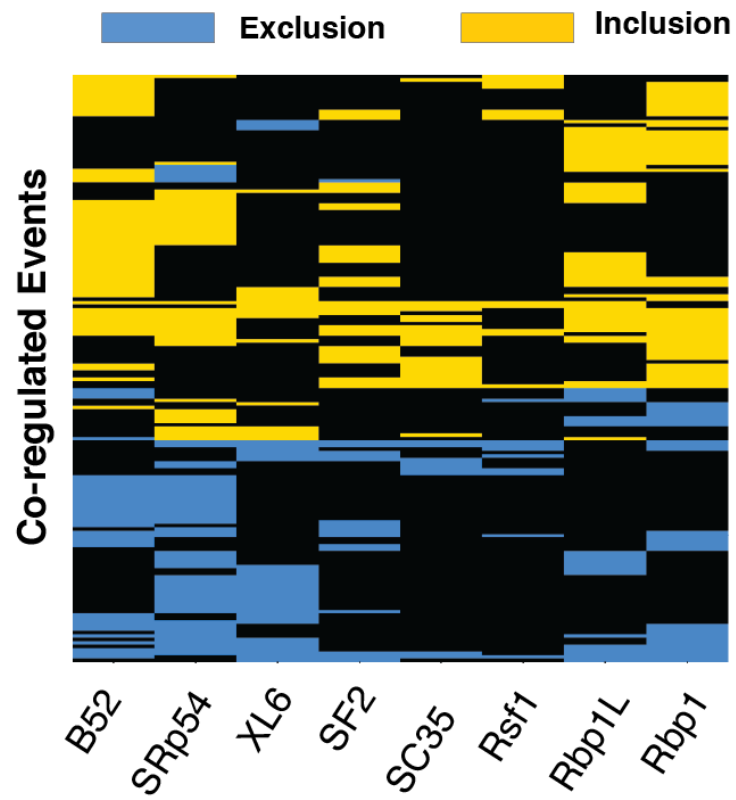


Figure 2.9: Multiple SR proteins co-regulate AS events. (A) Proportion of AS events that are uniquely (dark gray bar) or co-regulated by one or more other SR protein family member (red bar). (B) Heatmap of co-regulated events divided by events that have increased inclusion (yellow) and increased exclusion (blue) when a SR protein is depleted.

Su(Z)12 exhibits less inclusion when either SC35 or B52 protein levels are reduced, and a further decrease in inclusion is observed when both protein levels are reduced at the same time (Figure 2.11). Similarly, when XL6 and B52 are acting as a repressor of the cassette exon in *CG31638*, an increase in the magnitude of repression is observed when both proteins are knocked-down simultaneously (Figure 2.11). Surprisingly, this effect is even observed when the two SR proteins have antagonistic functions. XL6 acts as an activator, and B52 acts as a repressor, of the cassette exon in the *Dre4* transcript. When both XL6 and B52 are knocked-down together, the result is a splicing change that is near wild-type levels (Figure 2.11). In effect, the opposing regulation by these two SR proteins cancels each other out. These results indicate that the SR proteins co-regulate a majority of their AS targets, and this co-regulation is not redundant, but combinatorial.

2.3.4 SR proteins regulate promoter selection and 3' end processing of target transcripts

Many eukaryotic genes contain multiple promoters that determine a different start site and first exon of a transcript. Alternative promoter usage allows tissue-specific regulation of gene expression, and allows the target transcripts expression to be modulated at the transcriptional level. Alternative promoter usage can change the coding sequences of a

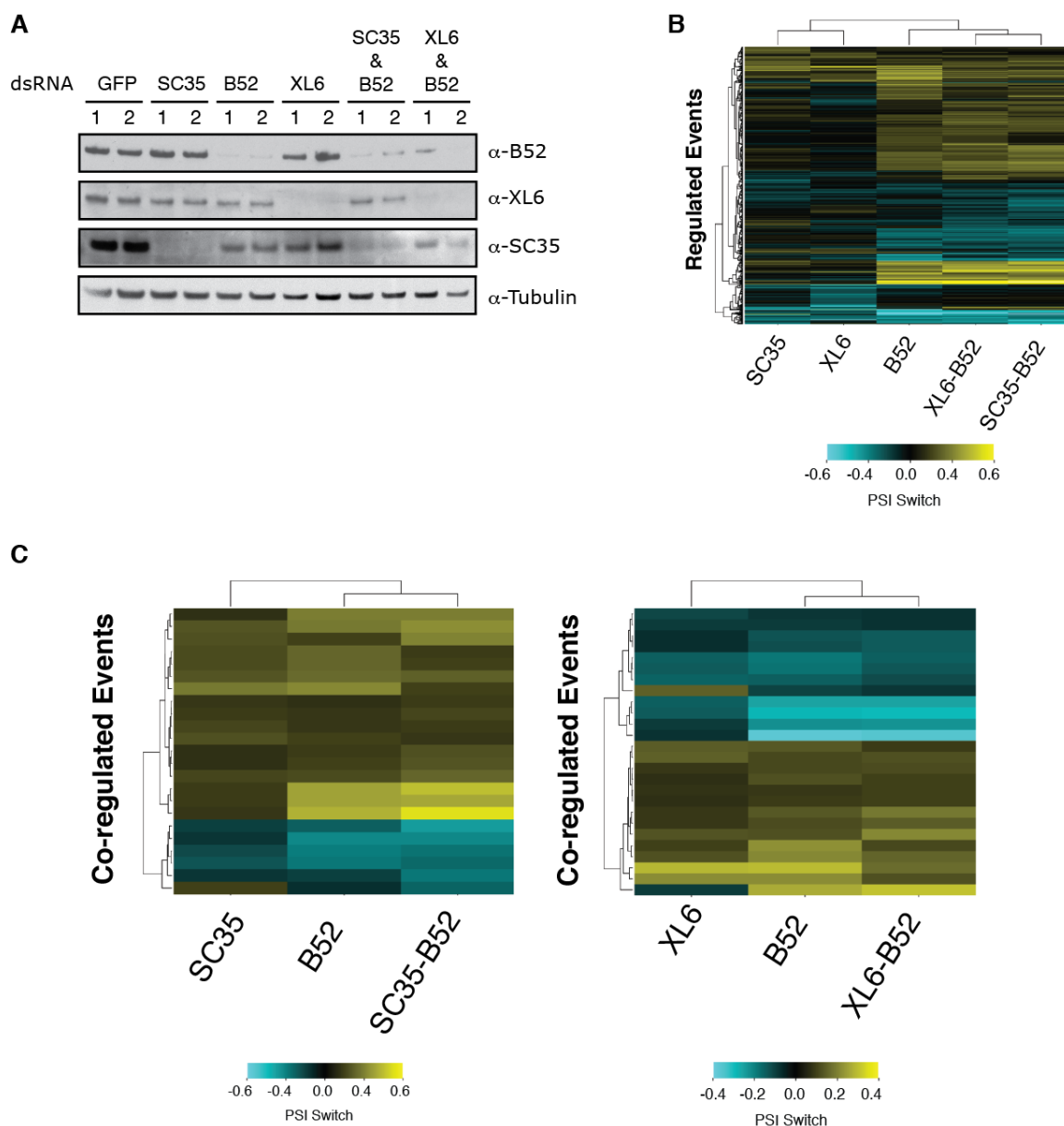


Figure 2.10: Combinatorial regulation of AS by the SR proteins. (A) Western-blot analysis demonstrating that individual and pair-wise SR protein expression could be reduced, specifically, in S2 cells. aWestern-blot analysis demonstrating (B) Heatmap shows hierarchical clustering of PSI-switch values for AS events changed with the depletion of SR proteins individually or pair-wise. (C) Heatmap shows hierarchical clustering of PSI-switch values for AS events co-regulated by SC35 and B52 (left panel) or XL6 and B52 (right panel) and the magnitude of splicing change when the pairs are simultaneously depleted.

protein, and produce functionally distinct protein isoforms. It can also change the regulatory regions in the 5' UTR that can affect protein expression levels and stability.

Using our alternative splicing event analysis from the mRNA-Seq data, we can monitor changes in alternative first exon usage. Surprisingly, SR proteins affect the frequency of usage of many alternative first exons (Figure 2.12A). Rbp1L regulates the usage of the fewest (71 first exons in 41 genes), and SRp54 regulates the most (278 first exons in 133 genes) first exons using our stringent AS filter ($p \leq 0.05$; $\Delta\text{PSI} \geq 0.1$). For example, when XL6 or B52 are reduced there is decreased usage of the distal promoter of the transcription factor *Nfat*, similarly, we found that there is increased usage of the two proximal promoters of the transporter gene *Indy* (Figure 2.12B). Currently, it is not clear how, mechanistically, SR proteins control promoter selection. A plausible mechanism could originate from the observation that SR proteins have been shown to associate with components of the transcriptional machinery, including the Pol II CTD. These protein-protein interactions could mediate the selection of the first exon of a transcript, independent of the SR protein's role in AS.

In addition to regulating the location of transcript initiation, SR proteins can also regulate where a transcript terminates, or alternative 3' processing. Recent studies have demonstrated a dynamic interplay between the cleavage/polyadenylation machineries and splicing regulators, and this results in altered protein products and transcript regulation. In some cases the alternative polyadenylation sites are located within the exons and introns in the protein-coding regions, termed coding region alternative polyadenylation (CR-APA), resulting in the production of different protein isoforms. We observed significant changes in the use of alternative poly (A) sites when SR proteins were

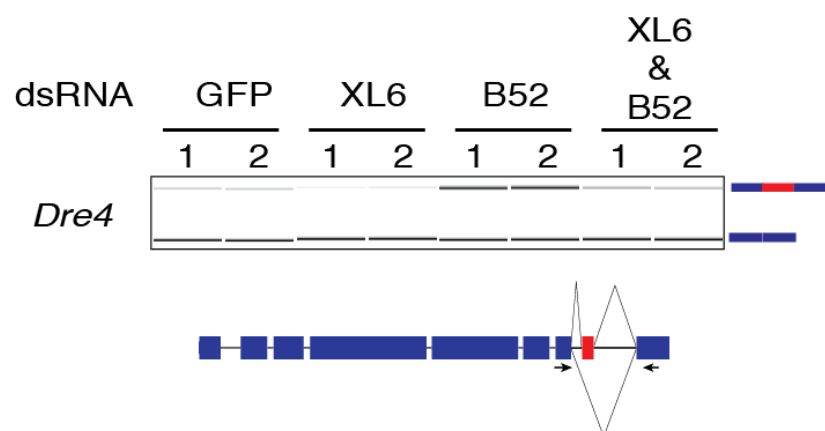
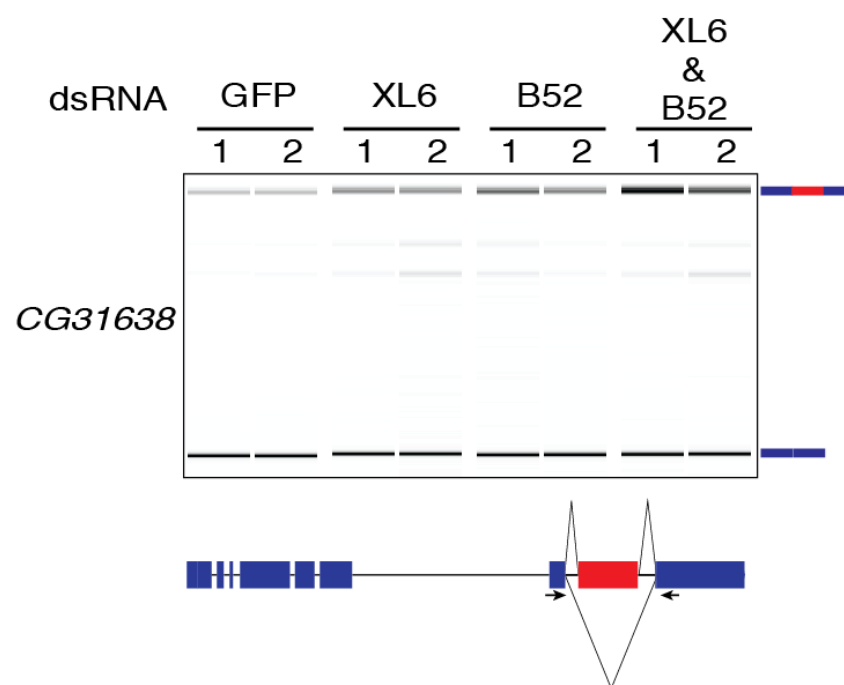
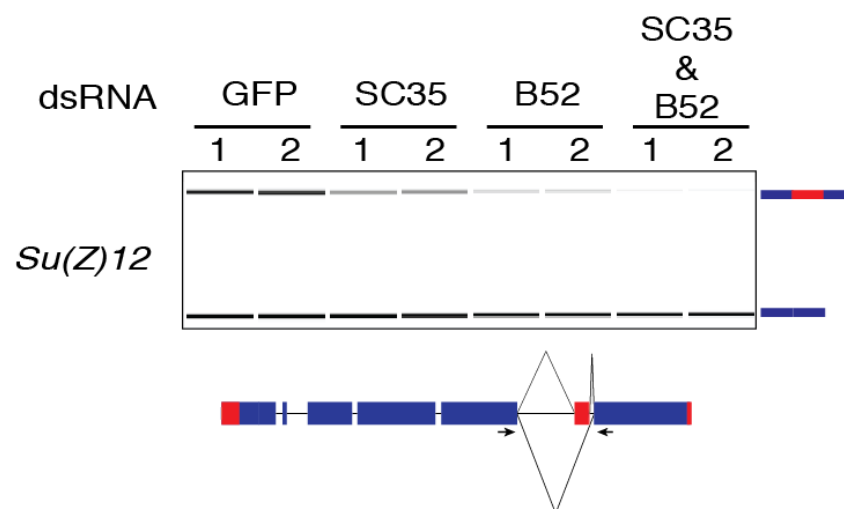


Figure 2.11: RT-PCR demonstrates combinatorial splicing control of specific SR targets. RT-PCR of co-regulated AS events affect in the same direction (cooperative; *Su(z)12* and *CG31638*) or opposing direction (antagonistic; *Dre4*) in conjunction with the simultaneous knock-down.

depleted (Figure 2.13A). Again, using our stringent filter ($p \geq 0.05$; $\Delta\text{PSI} \geq 0.1$), XL6 regulates the largest number of CR-APA events (53 events), and Rbp1 regulates the smallest number (6 events). The number of CR-APA events regulated by individual SR proteins are consistent with the number of other types of AS regulated by the SR proteins, indicating that this is a significant mechanism of regulation by the SR protein family. With the exception of SC35, the majority of regulated events shift to use of the proximal poly (A) site when SR protein level is reduced (Figure 2.13A). This means that in a majority of cases the SR proteins ensure that the longer transcript is produced. This could ensure inclusion of important regulatory sequences, and production of a functional protein.

An example of this type of regulation is found in the *Pnuts* gene, where two poly (A) sites are used to encode a transcription regulator (Figure 2.13B). The use of the proximal poly (A) site results in a transcript that is over 2000 nucleotides shorter in length, and a protein that is half as large than when the distal poly (A) site is used. Reduction of XL6 or B52 results in a decrease in use of the proximal, and an increase in use of the distal poly (A) site of this transcript (Figure 2.13B). Use of the proximal poly (A) site of *CG34439* results in a 74 amino acid protein compared to a 123 amino acid long protein when the distal site is used (Figure 2.13B). There is a reduction in use of the distal site

A

SR Protein	# of Regulated First Exons	# of Genes
B52	203	81
SRp54	278	133
XL6	296	120
SF2	197	90
SC35	138	68
Rsf1	129	58
Rbp1L	71	41
Rbp1	83	35

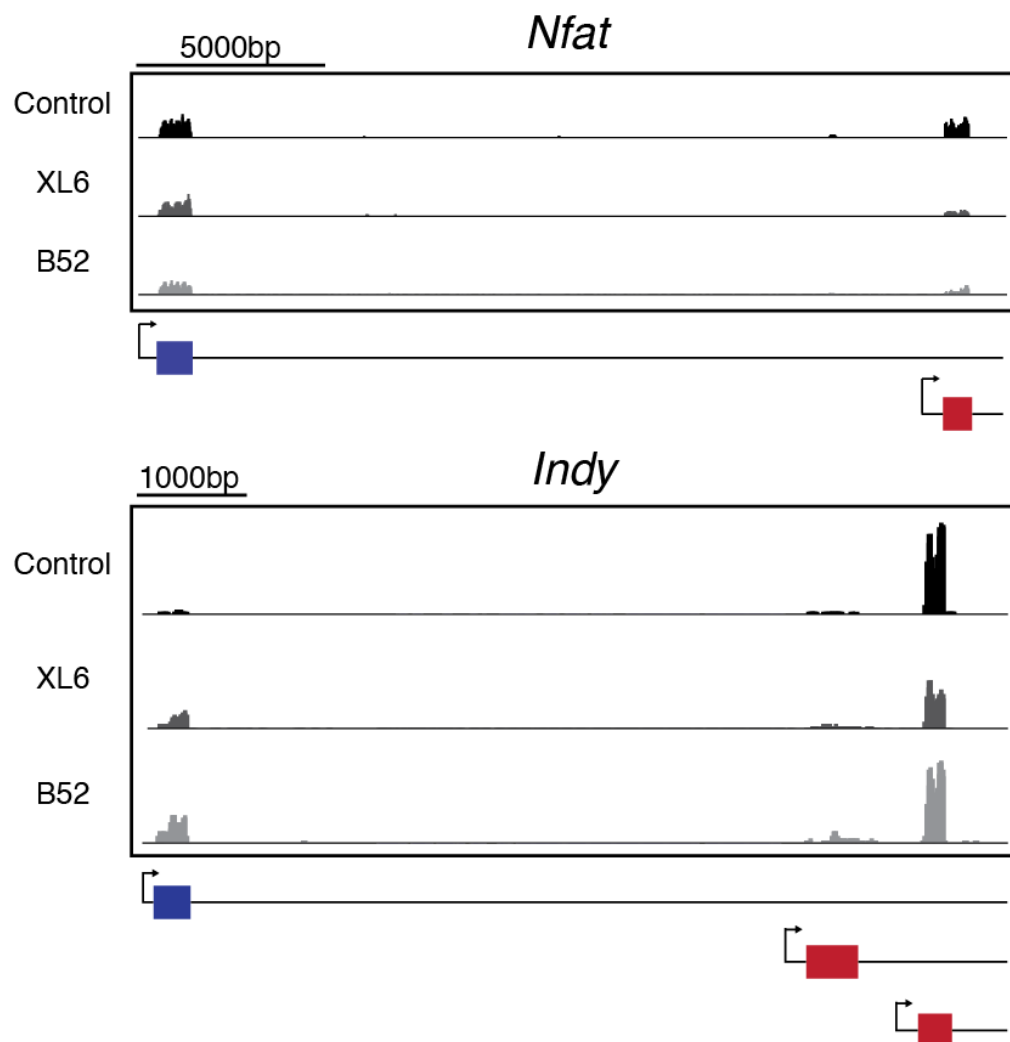
B

Figure 2.12: SR proteins regulate transcript start sites. (A) Number of transcripts with an SR regulated alternative 5' end, and total number of SR regulated first exons. **(B)** Examples of transcripts with altered 5' starts when SR proteins are depleted. Alternative first exons are denoted with red boxes, and the sequence tag density from RNA-Seq for control (black), XL6 (dark gray), and B52 (light gray) are above the gene model in the black box.

when XL6 or B52 protein levels are knocked-down, indicating that those two SR proteins are important for ensuring production of the longer protein isoform.

In addition to APA sites that occur in the protein-coding region, there are APA sites located in the 3' UTR that do not change the protein sequence, but can affect the expression of the protein, termed untranslated region alternative polyadenylation (UTR-APA). For example, extending the 3' UTR can introduce microRNA (miRNA) binding sites, or other regulatory sequences such as AU-rich elements (AREs) and RNA localization signals. SR proteins can also affect the use of poly (A) sites within the 3' UTR (UTR-APA) (Figure 2.14). Rbp1 regulates the smallest number (4 events) and XL6 regulates the largest (65 events) (Figure 2.14A). Similar to CR-APA, there is a shift to use of the proximal poly (A) site in a majority of cases when individual SR proteins are depleted, indicating a global role for SR proteins in ensuring that longer transcripts are produced (Figure 2.14A). Two notable exceptions are B52 and SC35, which in both CR-APA and UTR-APA, many more events result in use of the distal poly (A) site, indicating these SR proteins may promote production of shorter transcripts. Notable examples include *FabI*, a kinase that has been characterized as important for controlling cell growth and organ size (Figure 2.14B). When B52 protein levels are reduced there is a significant increase in use of the distal poly (A) site. Therefore, B52 is important to

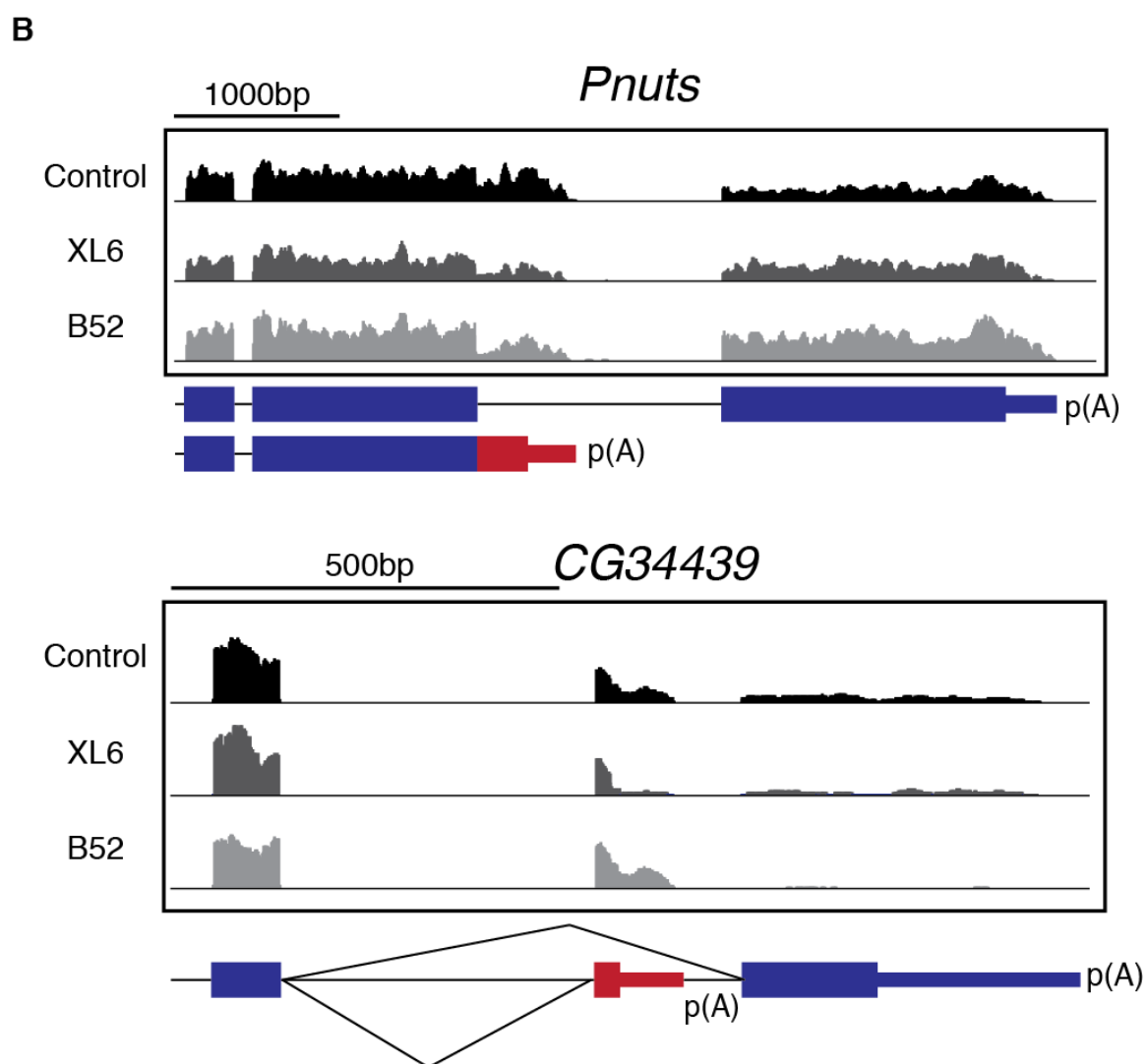
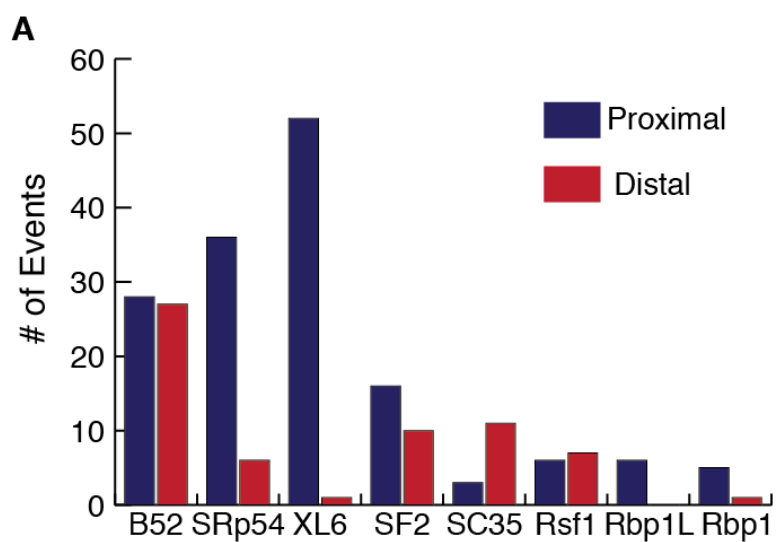


Figure 2.13: SR proteins regulate alternative polyadenylation that changes the protein coding sequence. (A) Total number of transcripts with APA within the coding region divided in shifts to the proximal (blue bar) or distal (red bar) polyadenylation site when an individual SR protein is depleted. **(B)** Examples of transcripts with APA when SR proteins are depleted. Alternative regions and poly(A) sites are denoted with red boxes, and the skinny portion denotes the 3' UTR. The sequence tag density from RNA-Seq for control (black), XL6 (dark gray), and B52 (light gray) are above the gene model in the black box.

promote the use of the shorter 3'UTR that excludes 1000 nucleotides of potential regulatory sequences. Similarly, *CG6700* is a gene that is highly expressed in the *Drosophila* brain. When XL6 is reduced there is a switch to use of the proximal poly (A) site (Figure 2.14B). Therefore, XL6 is responsible for maintaining the long 3' UTR of this transcript. The differences in proximal or distal poly (A) site regulation by the SR proteins could be a result of cell or tissue specific expression regulation.

One striking difference between CR-APA and UTR-APA is the mechanism that may be used by SR proteins to regulate APA site choice. CR-APA could be the result of differential selection of splice sites, by the SR proteins, that include or remove poly (A) selection regulatory sequences from the transcript, and ultimately change where the transcript is cleaved and polyadenylated. However, UTR-APA involves differential selection of poly (A) sites, independent of splice site selection (Figure 2.14B). Many RNA-binding proteins have been implicated in UTR-APA control by enhancing or antagonizing the action of the formation of the 3' processing complex, but SR proteins have only been implicated in CR-APA [128]. These results extend regulation of UTR-APA to the SR protein family, and also suggest SR protein interaction with the 3' processing machineries.

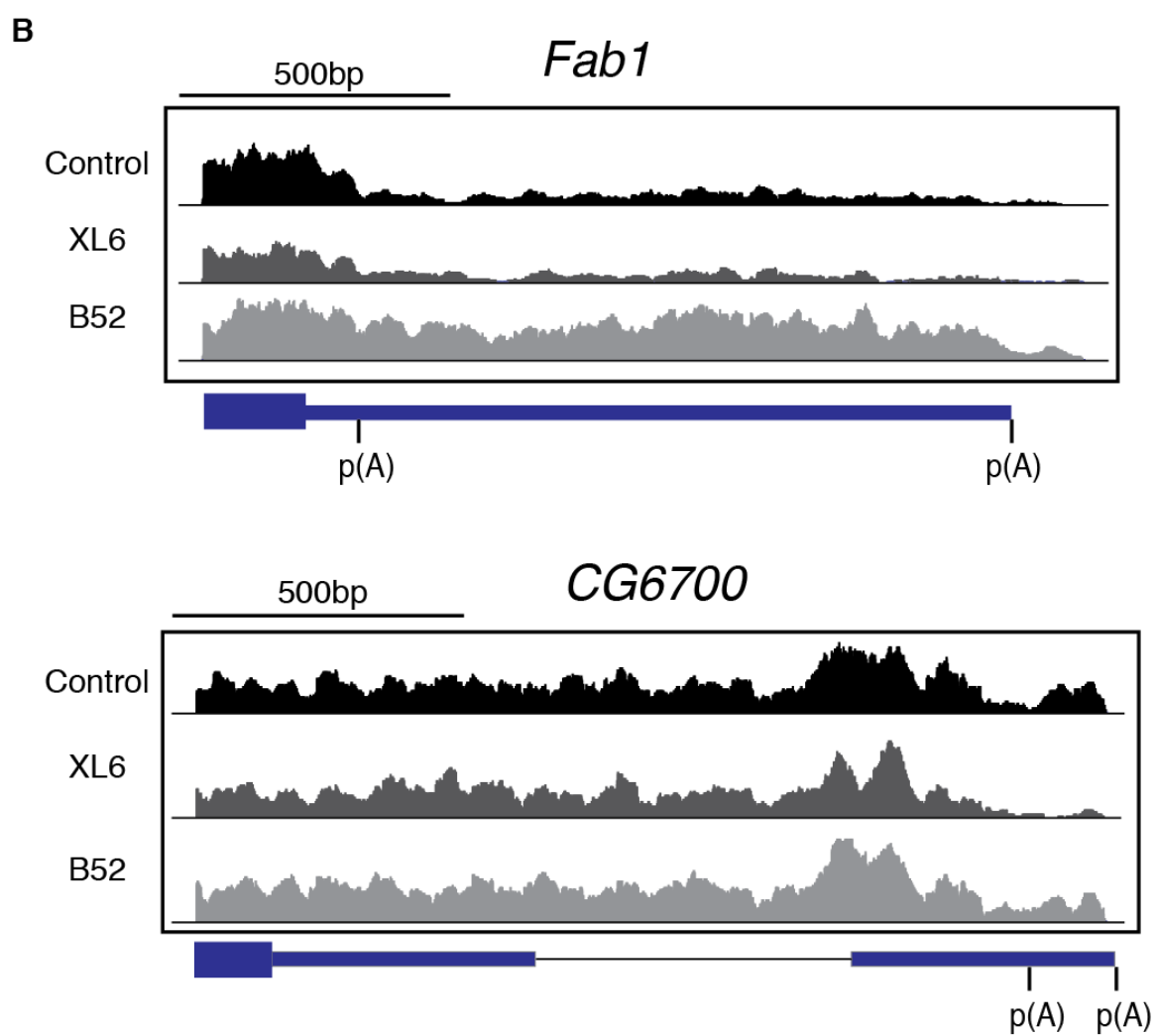
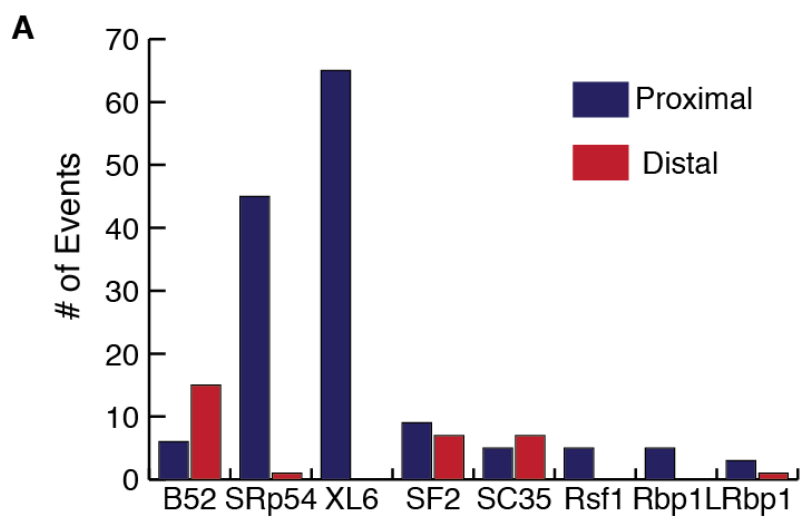


Figure 2.14: SR proteins regulate alternative polyadenylation in the 3' UTR independent of AS. (A) Total number of transcripts with APA within the 3' UTR divided in shifts to the proximal (blue bar) or distal (red bar) polyadenylation site when an individual SR protein is depleted. **(B)** Examples of transcripts with APA when SR proteins are depleted. Alternative polyadenylation sites are denoted by p(A), and the skinny portion denotes the 3' UTR. The sequence tag density from RNA-Seq for control (black), XL6 (dark gray), and B52 (light gray) are above the gene model in the black box.

2.3.5 SR proteins are required for maintaining the cellular transcriptional profile

In addition to regulating transcript length and sequence composition, SR proteins can modulate gene expression levels at other stages of RNA metabolism, such as mRNA export, stability, and translation. SR proteins have also been shown to affect transcriptional elongation, and associate with sites of active chromatin [72, 129]. To identify transcripts that had significant changes in their levels of expression when an individual SR protein is absent, we computed the number of uniquely mapped reads per kilobase of exon, to normalize for length of transcripts, per million mapped reads (RPKM), for each gene in the SR protein RNAi RNA-Seq samples. Next, we used the edgeR package to calculate significant differential gene expression levels between control and SR-depleted RNA-Seq samples. Surprisingly, there are a large number of genes whose expression is significantly changed (≥ 2 Fold-Change) when an individual SR protein is depleted (Figure 2.15A). 539 genes have altered expression levels when XL6 is depleted (the lowest), and 1,230 genes when B52 is depleted (the highest). This represents between 6.1% and 13.8% of reliably detected genes expressed in S2 cells (8,891 genes). In total, SR proteins regulate the expression of 2,283 (1,661 up-regulated and 722 down-regulated) genes. This means that SR proteins can regulate the expression of 25.7% of the expressed genes in S2 cells, a striking number. Expression levels are

changed in both directions (increased and decreased), but a majority of targets are up-regulated when an individual SR protein is reduced (Figure 2.15A).

Hierarchical clustering of the affected genes in each experiment reveals that, similar to splicing regulation, SR proteins affect the gene expression of a distinct and similar set of genes (Figure 2.15B). GO analysis of significantly enriched GO terms, by TopGO (Top 20 nodes), revealed that transcripts that increase in expression level in the absence of SR proteins play key roles in cell growth and signaling, and transcripts whose expression level is decreased are important for reacting to external stimuli and tissue development (Figure 2.16). Therefore, altering SR protein levels can assist the cell in the rapid response to different cellular cues.

The number of genes with differential expression upon SR protein knock-down exceeds the total number of genes identified with changes in simple AS events. This indicates that the SR proteins may influence the expression of their targets independent of their role in AS. To investigate if SR protein AS regulation plays a major role in controlling overall transcript levels, we compared the previously identified regulated AS events, and determined if there were significant changes in gene expression levels. In a vast majority of cases (> 62.5% for Rbp1 and > 80% for all other SR proteins) genes affected at the level of AS had no significant changes in transcript abundance (Figure 2.17).

This indicates that regulation of transcript expression by the SR proteins is downstream of direct AS targets, or is independently regulated by a mechanism other than AS. GO analysis of AS regulated targets revealed genes involved in cell signaling and development. This could mean that by altering the splicing of a transcriptional or signal-

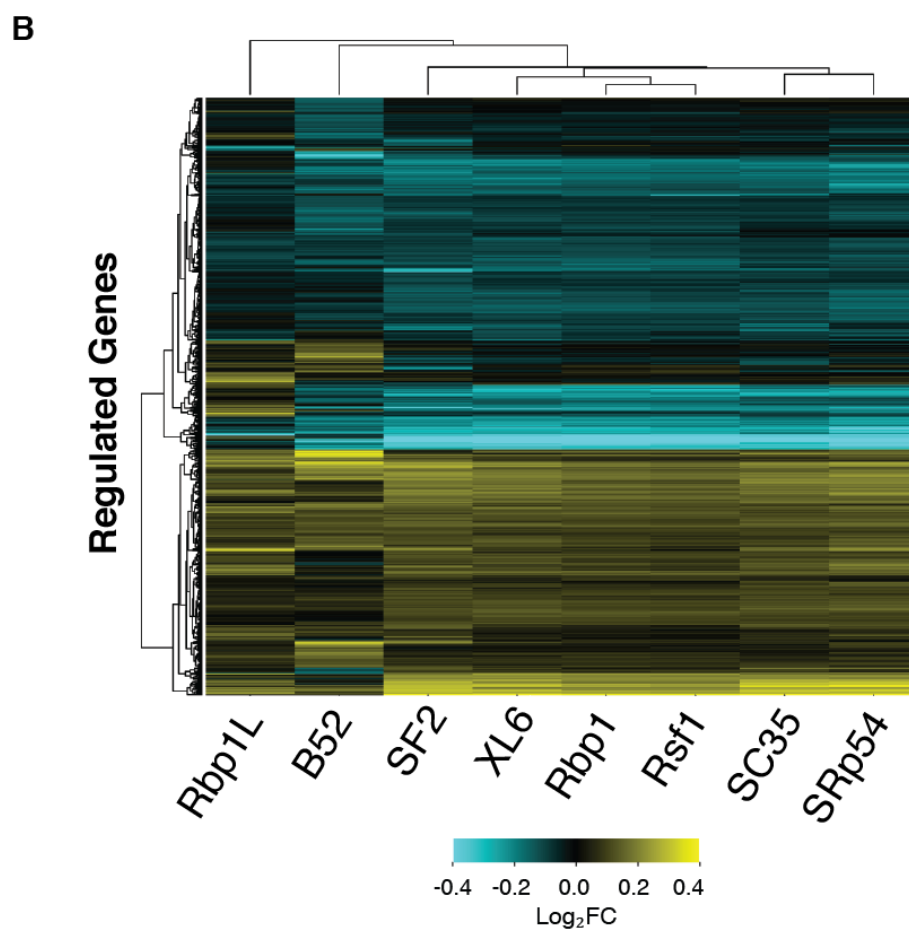
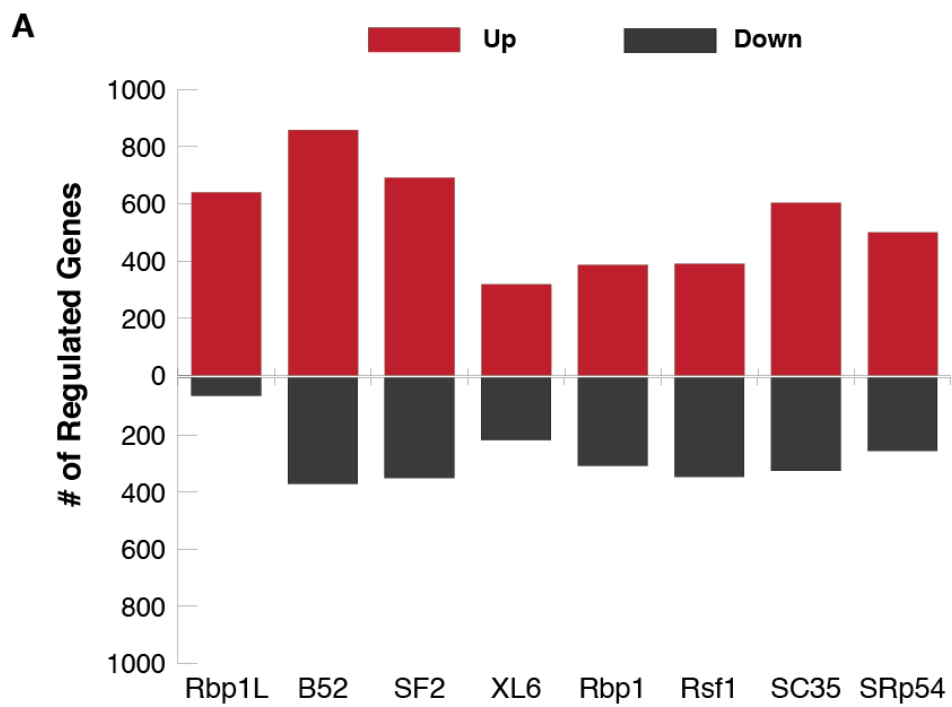


Figure 2.15: SR proteins regulate RNA transcript levels. (A) Total number of genes with differential expression in each SR protein knock-down experiment (increased expression, red; decreased expression, dark gray). **(B)** Heatmap showing hierarchical clustering of calculated fold-change (\log_2) values for genes with altered transcript levels when an SR protein is depleted.

transduction pathway regulator could contribute to widespread changes in downstream gene expression. Indeed, SR proteins regulate the AS of key transcription factors (*Su(z)12*, *Iswif*, *Hsf*, *Skd*) identified by our study. Alternatively, SR proteins may regulate gene expression of targets by directly affecting their transcription and/or stability. Either way, this provides a sensitive and rapid mechanism to regulate an organism's gene expression profile.

2.4 Discussion

In this chapter, we identified global targets of AS regulation for each member of the SR protein family in *Drosophila*. SR proteins share a similar protein-domain structure and organization, and since multiple members could complement splicing-deficient S100 extract, SR proteins were thought to be functionally redundant [130]. The observations that (i) when deletion of an individual SR protein resulted in embryonic lethality and (ii) that the phenotype could not be rescued by over-expression of another SR protein family member demonstrated that SR proteins had specific functions and were not entirely redundant. More recently, it was discovered that individual SR proteins have distinct effects on AS outcomes. While many members of the SR protein family have been implicated in AS regulation, there has never been a comprehensive study to determine if all SR proteins regulate AS and how this regulation is achieved. Many of the previous

Gene ontology analysis of genes up-regulated when SR proteins are depleted in S2 cells

GO.ID	Term	Fisher.Test
GO:0006811	ion transport	1.70E-06
GO:0055085	transmembrane transport	2.10E-06
GO:0007156	homophilic cell adhesion	1.80E-05
GO:0006030	chitin metabolic process	6.00E-05
GO:0006096	glycolysis	7.30E-05
GO:0055114	oxidation-reduction process	0.00011
GO:0016337	cell-cell adhesion	0.00013
GO:0007155	cell adhesion	0.00016
GO:0022610	biological adhesion	0.00018
GO:0035151	regulation of tube size, open tracheal s...	0.00037
GO:0042044	fluid transport	0.00092
GO:0035150	regulation of tube size	0.00095
GO:0006040	amino sugar metabolic process	0.00097
GO:1901071	glucosamine-containing compound metaboli...	0.00097
GO:0007354	zygotic determination of anterior/poster...	0.00104
GO:0016052	carbohydrate catabolic process	0.00106
GO:0030001	metal ion transport	0.00137
GO:0006812	cation transport	0.00141
GO:0006022	aminoglycan metabolic process	0.00214
GO:0009636	response to toxin	0.00231

Gene ontology analysis of genes down-regulated when SR proteins are depleted in S2 cells

GO.ID	Term	Fisher.Test
GO:0050830	defense response to Gram-positive bacter...	9.80E-09
GO:0009617	response to bacterium	1.20E-07
GO:0042742	defense response to bacterium	1.30E-07
GO:0006952	defense response	4.40E-07
GO:0051704	multi-organism process	1.20E-06
GO:0019731	antibacterial humoral response	1.20E-05
GO:0051707	response to other organism	1.30E-05
GO:0009607	response to biotic stimulus	1.80E-05
GO:0000245	spliceosomal complex assembly	0.00015
GO:0050829	defense response to Gram-negative bacter...	0.00021
GO:0036293	response to decreased oxygen levels	0.00063
GO:0070482	response to oxygen levels	0.00063
GO:0006508	proteolysis	0.00108
GO:0006955	immune response	0.0011
GO:0001666	response to hypoxia	0.00112
GO:0071704	organic substance metabolic process	0.00135
GO:0051310	metaphase plate congression	0.00145
GO:0007498	mesoderm development	0.00159
GO:0050000	chromosome localization	0.00203
GO:0051303	establishment of chromosome localization	0.00203

Figure 2.16: Over-represented GO-terms in SR protein gene expression regulated targets. GO-analysis of SR regulated genes that change greater than 2-fold in expression when an SR protein is depleted. Enriched GO-terms determined using TopGO.

studies on SR protein AS regulation used *in vitro* systems, or were based on single gene models. There has never been a systematic, genome-wide, analysis of AS for a majority of the SR protein family members. Recent advances in the ability to rapidly characterize AS have allowed for a more comprehensive analysis of AS. It is now possible to use RNA sequencing to identify genome-wide, endogenous, AS events regulated by individual SR proteins. A previous study, utilizing splicing-sensitive microarrays profiling genome-wide AS events regulated by *Drosophila* SR proteins, SF2 and B52, revealed that they shared a remarkable number of targets. Furthermore, the study pointed out that in addition to overlapping targets, SF2 and B52 individually regulate many more unique AS events [116]. This suggested that SR proteins could have overlapping functions, but also have discrete roles in AS regulation. Here we extend this observation of AS co-regulation to the entire family of SR proteins. Using RNA-Seq, we identified 561 simple AS events regulated by individual SR proteins, and reveal that while individual SR proteins can regulate unique targets, a majority of SR regulated AS events are controlled by multiple SR protein family members. We also noted the direction of the splicing change, and found that all SR proteins can function as alternative splicing event activators and silencers. In addition, we found that SR proteins act in concert (regulation is in the same direction) to regulate overlapping targets a vast majority of the time, but there are a few AS events that are regulated antagonistically by multiple SR proteins. Although, we observed no pair-wise SR protein patterns of regulation, it would be

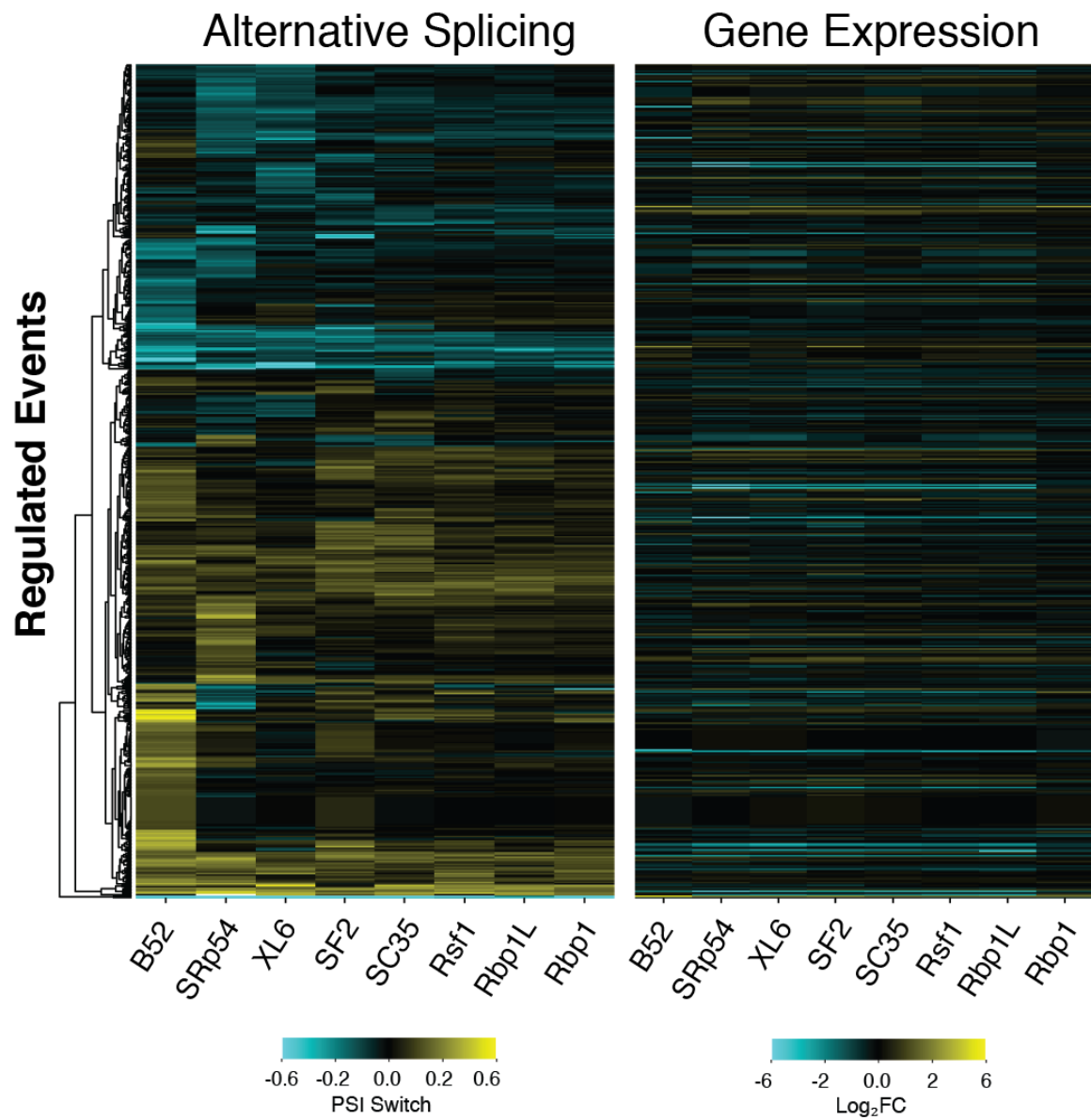


Figure 2.17: SR protein regulation of gene expression is independent of AS regulation. Heatmaps comparing identified AS regulated events (left heatmap) and their corresponding gene expression changes (right heatmap).

interesting to determine features of unique and co-regulated SR protein targets that would explain why some AS events are only regulated by one member of the SR protein family.

The observation of cooperative regulation of AS has been demonstrated for the hnRNP family of splicing regulators, in both *Drosophila* and mammals [74, 112]. However, the mechanism of the cooperative regulation has not been uncovered. To explore if the SR proteins co-regulate AS in a combinatorial manner we depleted two SR proteins, simultaneously. We showed that the observed splicing change increased in magnitude when two SR proteins were knocked-down together for a large proportion of co-regulated targets. Surprisingly, this effect is even observed on AS events that are regulated antagonistically. This observation reveals insights governing the co-regulation of AS by the SR proteins, and shows that SR protein co-regulation is achieved by combinatorial regulation.

In addition to AS of the same primary transcript, different mRNA isoforms can be initiated at alternative promoters, and by differential termination/processing of the 3' end. We found that SR proteins affect the frequency of usage of many alternative first exons, and regulate the 3' end of a variety of transcripts. In both cases this regulation can occur independent of the SR protein's role in splice site selection. Recent studies have demonstrated a dynamic interplay between splicing regulators, the cleavage/polyadenylation machineries, and components of polymerase II [120]. The observation that SR proteins can regulate alternative promoters, and alternative polyadenylation, expand SR protein interaction beyond the spliceosome. One striking difference between canonical AS changes, and differences in a transcript's start and end, is that many changes at the beginning and end of a transcript do not change the protein

coding sequence. Rather, it modifies the 5' and 3' UTR that is used for regulation of downstream RNA processing events, such as RNA localization and translation [119, 121, 122]. SR proteins have been implicated in the regulation of many of these downstream RNA processing events. Perhaps one mechanism of this regulation is through interaction with the transcription and polyadenylation machineries to regulate the length of a transcripts 5' and 3' UTR. It would also be interesting to see if translation or RNA localization of these transcripts is affected upon SR protein depletion. On the contrary, it would be interesting to determine if components of the 5' and 3' processing machineries can influence SR protein function in AS.

Mounting evidence suggests that transcription and AS are coupled to one another. The CTD of Pol II has been found to be associated with various RNA processing factors, including SR proteins [124]. In addition, SR proteins have been shown to associate with active chromatin, and affect transcriptional elongation [72]. This suggests that SR proteins may have an early function in gene expression regulation, possibly upstream of AS regulation. Indeed, there are a large number of genes whose expression is changed when an individual SR protein is depleted. Similar to AS regulation, SR proteins affect the expression of a very similar, overlapping, set of targets. We also found that most AS targets did not have changes in gene expression, and the number of genes with differential expression exceeds the total number of genes with changes in AS. This indicates that SR proteins may influence the expression of their targets independent or downstream of their role in AS regulation. While we observe SR protein regulation of the AS of transcriptional regulators, further studies will need to be done to elucidate if gene expression regulation occurs downstream of the AS decision. Regardless, our study

reveals a large number of genes whose expression relies on SR protein levels. This may be a novel mechanism for an organism to regulate its gene expression profile, rapidly, by modulating the expression of one or more SR proteins.

Chapter III: SR Protein Domain Contribution to Alternative Splicing Regulation

3.1 Introduction

SR proteins are defined by their ability to promote constitutive and alternative splicing, but are also grouped based on similar protein domain architecture (Figure 3.1). SR proteins consist of one or two N-terminal RNA recognition motifs (RRM) and a C-terminal domain enriched in arginine and serine (RS-domain). Typically, it has been thought that SR proteins bind pre-mRNA via their RRM domains, and their RS-domains function in protein-protein interactions with key components of the basal splicing machinery [131]. Using yeast two-hybrid assays, ASF/SF2 was shown to interact with the U1 and U2AF snRNPs to promote recruitment of these complexes to splice sites. This interaction required the RS-domains of each protein [56, 132]. One drawback of these *in vitro* assays is that it has yet to be demonstrated that these protein-protein interactions occur in a functional spliceosome.

SR proteins have also been reported to promote splicing through RS-domain:pre-mRNA interactions. SR protein RS-domains bound to ESEs strengthened base-pairing of U5 and U6 snRNAs through RS-domain contact with the branch-point, and promote splicing [133]. On the other hand, the addition of free RS-domain to nuclear extract can rescue the splicing of a *doublesex* splicing substrate, arguing that splicing can be accomplished without a pre-mRNA binding domain [134]. Additionally, it has been shown that ASF/SF2 can interact with U1-70K snRNP through RRM-domain interactions in early spliceosome assembly, indicating that the RRM-domain can also contribute to protein-

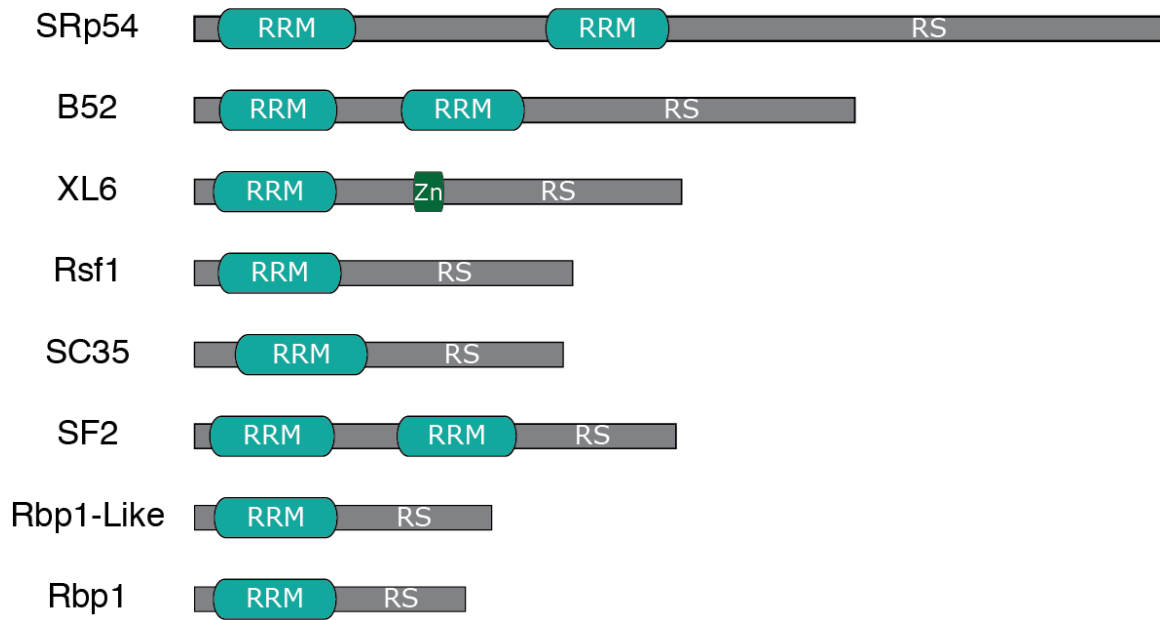


Figure 3.1: SR protein domain organization. The eight identified SR proteins in *Drosophila melanogaster* organized, descending, by RS-domain length.

protein interactions. [135]. It still remains unclear the contribution of each domain during splicing *in vivo*.

Similar to the RRM-domain, the content and the length of the SR protein RS-domains are conserved from *Drosophila* to humans, and are targeted for reversible phosphorylation at serine residues by multiple kinases. The phosphorylation state of the RS-domain can affect SR protein localization, RNA-binding, and protein-protein interactions [65, 93, 136]. Using *in vitro* splicing assays, SR proteins lacking RS-domains were not able to complement S100 splicing extracts, but recently, it was demonstrated that the RS-domain of ASF/SF2 is not required for the splicing of several, but not all, pre-mRNAs. This reliance on the RS-domain for splicing regulation was found to be related to the strength of the 3' splice site [137]. SR proteins differ in their abilities to complement splicing

deficient S100 extracts. Additionally, different RS-domains have been shown to differ in their potency when tethered to the same ESE. Furthermore, by increasing the number of binding sites, or length of the RS-domain, additively increased the efficiency of splicing [60]. This raised the possibility that differences in splicing activity between SR proteins could be attributed to the strengths of the RS-domains in addition to the specificity of RNA binding.

Despite careful dissection of SR protein domain contribution to splicing, *in vitro*, there has never been a study that examines the contribution of each domain to regulation of endogenous targets. Additionally, previous studies have focused on the role for SR protein domains in constitutive splicing, and have never examined their requirement in AS. In this chapter, we set out to identify the SR protein domains required, *in vivo*, for regulation of AS targets. We first demonstrate that an RS-domain is required for proper regulation of endogenous AS targets. Next, we show that the specificity and direction of AS regulation is dictated by the RRM-domain. Although, another SR protein's RS-domain can functionally substitute, the specificity and direction of splicing change relies on the RRM-domain.

3.2 Materials and Methods

3.2.1 Construction of 2x Flag-HA XL6 WT and Mutant constructs

Full length XL6, XL6ARS, and XL6-B52RS were PCR amplified and cloned into pDONR221 by BP reaction (Gateway cloning; Invitrogen). XL6-B52RS was amplified using a two-step PCR. First, the RRM-domain of XL6 and RS-domain of B52 were

amplified, separately. Next, the purified PCR products were mixed together and amplified with outside primers to generate the fusion.

XL6_FL_SP

5- GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGTCGCGCCATCCGAGC-3

XL6_FL_ASP

5- GGGGACCACTTTGTACAAGAAAGCTGGGTAGTCCCTTGAAACGGA-3

XL6ΔRS_ASP

5-

GGGGACCACTTTGTACAAGAAAGCTGGGTAACACGGGCTCGGCGCCCGCAA
A-3

XL6-B52RS_SP

5- GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGTCGCGCCATCCGAGC-3

XL6_RRM_ASP

5- CTCCGCGACGATCCTCGACTTCGCGACAGTGGCGAGCGAA-3

B52_RS_SP

5- TTCGCTCGCCACTGTCGCGAAGTCGAGGATCGTCGCGGAG-3

XL6-B52RS_ASP

5- GGGGACCACTTTGTACAAGAAAGCTGGGTAATCGTCCATGCTCTCGTT-3

Next, the BP reaction was transformed into DH5a cells, plated on kanamycin plates and screened for positive clones. The constructs were inserted into a destination vector that contained the metallothionein promoter and a C-terminal 2x Flag and 2x HA tag (pMT-C2FL2HA) by LR reaction (Gateway cloning; Invitrogen). The LR reaction was transformed into DH5a. The LR reaction was transformed into DH5.HA

F vector that contained the

3.2.2 Generation stably transfected cell lines

400ng of pMT-C2FL2HA plasmid DNA containing XL6, XL6ΔRS, and XL6-B52RS and 40ng of plasmid carrying the gene for blasticidin resistance (pCO-BLAST) were transfected into 2.0×10^6 *Drosophila* S2 cells, seeded in a 6-well tissue culture dish, using Effectene transfection reagent (Qiagen) according to the manufactures protocol. After 48hrs, blasticidin (Invivogen) was added to the transfected cells at 25μg/ml.

Resistant cells were expanded, and maintained, in Schneiders media with 10%FBS and 25µg/ml Blasticidin.

3.2.3 Generation of dsRNA against XL6 5' and 3' UTRs

dsRNA was made against the 5' and 3' UTRs of XL6 to using the procedure described in section 2.2.2. Primers used to generate the T7-DNA template were:

T7_XL6_5UTR_S
5-TAATACGACTCACTATAGGGAGGACCAGACACTCCCGCCTTTTGC-3
T7_XL6_5UTR_AS
5-TAATACGACTCACTATAGGGAGGATGTTTATTTCGCTTTTCCCCG-3
T7_XL6_3UTR_S
5-TAATACGACTCACTATAGGGAGGATAGAAATGTATCGGCAGCCGC-3
T7_XL6_3UTR_AS
5-TAATACGACTCACTATAGGGAGGATAATTTTGGCATTTAATTTAT-3

3.2.4 Expression of WT and Mutant XL6 constructs in *Drosophila* S2 cells

1.0 x 10⁶ cells of each stably transfected cell line was seeded in 4 wells each of a 12-well tissue culture dish. 10µg of Control dsRNA or dsRNA against the UTRs of XL6 was added. After 48hrs, another boost of 10µg of dsRNA was added to the cells, and the expression of the tagged constructs was induced by adding CuSO₄ at a concentration of 100µM. The cells were then incubated for an additional 48hrs. The cells were harvested, and the RNA purified using the RNeasy kit following the manufacturer's protocol (Qiagen #74106).

3.2.5 Western-blot to determine construct expression and endogenous XL6 knock-down

After RNAi and expression of XL6 protein constructs, 10% of the cells were saved and 3% loaded on a 10% SDS-PAGE gel, transferred, and blotted with antibody raised against XL6 according to section 2.2.5. Antibody against H3 was used as a loading control.

3.2.6 RT-PCR of SR protein targets

5µg of purified RNA added to a randomly primed/oligo d(T)₁₅ reverse transcription reaction using the procedure described in section 2.2.8. PCR was performed as described in section 2.2.8 on AS events in *Dre4*, *Vps35*, *DppIII*, *CG12065*, and *Syb*. PCR was also performed on cDNA from an RNAi of B52, in wild-type S2 cells, to assay B52 contribution to regulation.

3.3 Results

It has been proposed that SR proteins can cooperate with spliceosomal components (such as U1 snRNP and U2AF) in recognition of the 5' or 3' splice sites, and this effect is most likely mediated by specific interactions between the RS-domains. Interestingly, *in vitro* splicing of several pre-mRNAs do not require an RS-domain at all, and this directly correlates with the strength of the 3' splice site [137]. Therefore, we set out to identify the SR protein domains required, *in vivo*, for regulation of endogenous AS targets.

First, we engineered stably transfected S2 cells that can be induced, using a metallothionein promoter, to express a wild-type Flag-HA tagged XL6 (XL6-FL) and

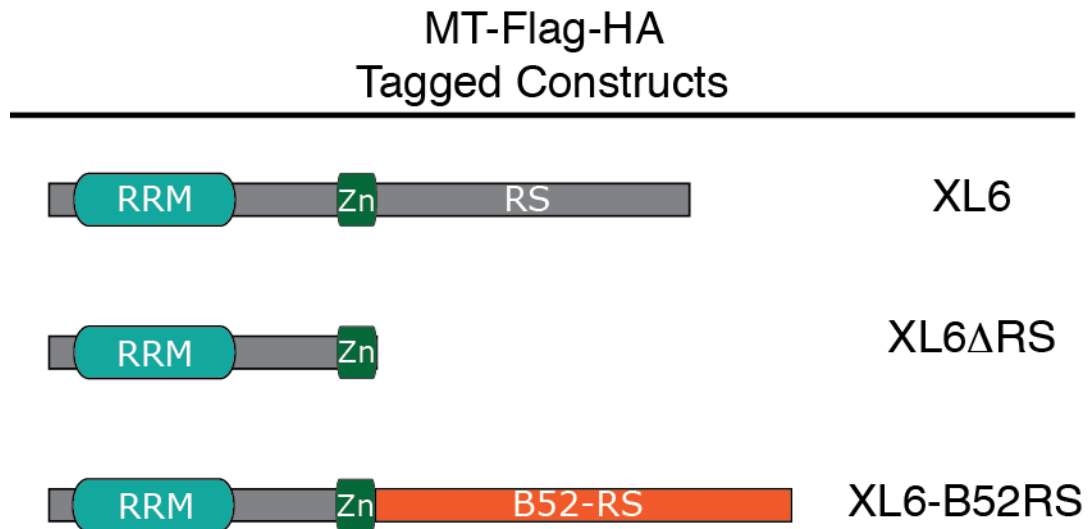


Figure 3.2: Diagram of WT XL6 and variant proteins. Diagram of wild-type XL6, XL6 Δ RS, and XL6-B52RS 2xFlag2xHA tagged proteins, showing relevant domains. The RNA binding domain (RRM; blue), zinc-knuckle (Zn; green) and the C-terminal arginine/serine domain (RS) organization of copper inducible constructs.

XL6 lacking the RS-domain (XL6 Δ RS), to examine the RS-domain requirement for AS regulation (Figure 3.2). RS-domains from different SR proteins differ in sequence composition and length, but have been demonstrated to be, functionally, substituted for one another *in vitro* (Figure 3.1). However, there are differences in potency of RS-domains when tethered to a similar ESE. We therefore designed a mutant that substitutes the RS-domain of B52 for the RS-domain of XL6 (XL6-B52RS) to test whether a different RS-domain can functionally substitute for antagonistic and specific XL6 AS targets.

Wild-type and mutant forms of XL6 were expressed in S2 cells, RNA was extracted, and subjected to RT-PCR. To eliminate effects from endogenous XL6, we reduced its expression by using dsRNA that targets the 5' and 3'UTR of the protein. This allows

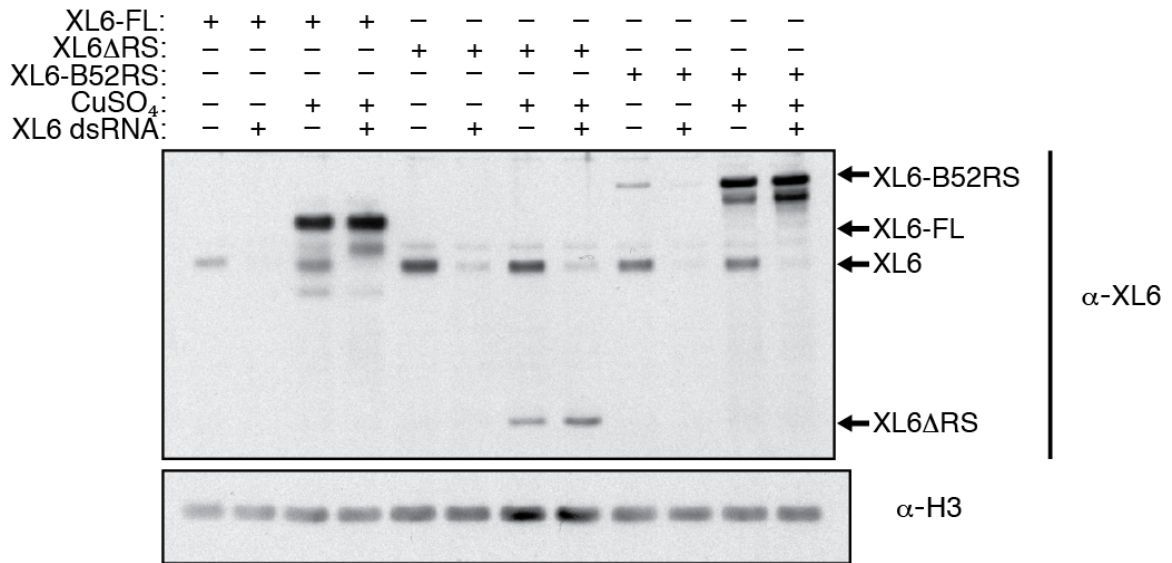


Figure 3.3: Expression of recombinant XL6 constructs in *Drosophila* S2 cells. Expression of endogenous and recombinant XL6 constructs in *Drosophila* S2 cells analyzed by western-blot. Expression of recombinant XL6 is assayed in the presence or absence of dsRNA against endogenous XL6. Histone H3 was used as a loading control.

reduction of endogenous XL6 without affecting the expression of the tagged XL6 constructs (Figure 3.3). First, we examined two AS events that are regulated by XL6 and B52 antagonistically (*Dre4* and *Vps35*). XL6 is an activator of the *Dre4* cassette exon and repressor of the *Vps35* competing donor event, while B52 represses *Dre4* and activates *Vps35* AS events (Figure 3.4A;compare lanes 1 and 2 and Figure 3.4B). Over-expression of wild-type XL6 enhances the activation and repression of *Dre4* and *Vps35* AS events, respectively. This enhancement remains even when endogenous XL6 protein levels are reduced (Figure 3.4A;compare lanes 3 and 4 to 1 and 2). This demonstrates that there is specific, concentration dependent, regulation of these events by XL6. It also shows that by over-expressing XL6 the repression by B52 can be overcome.

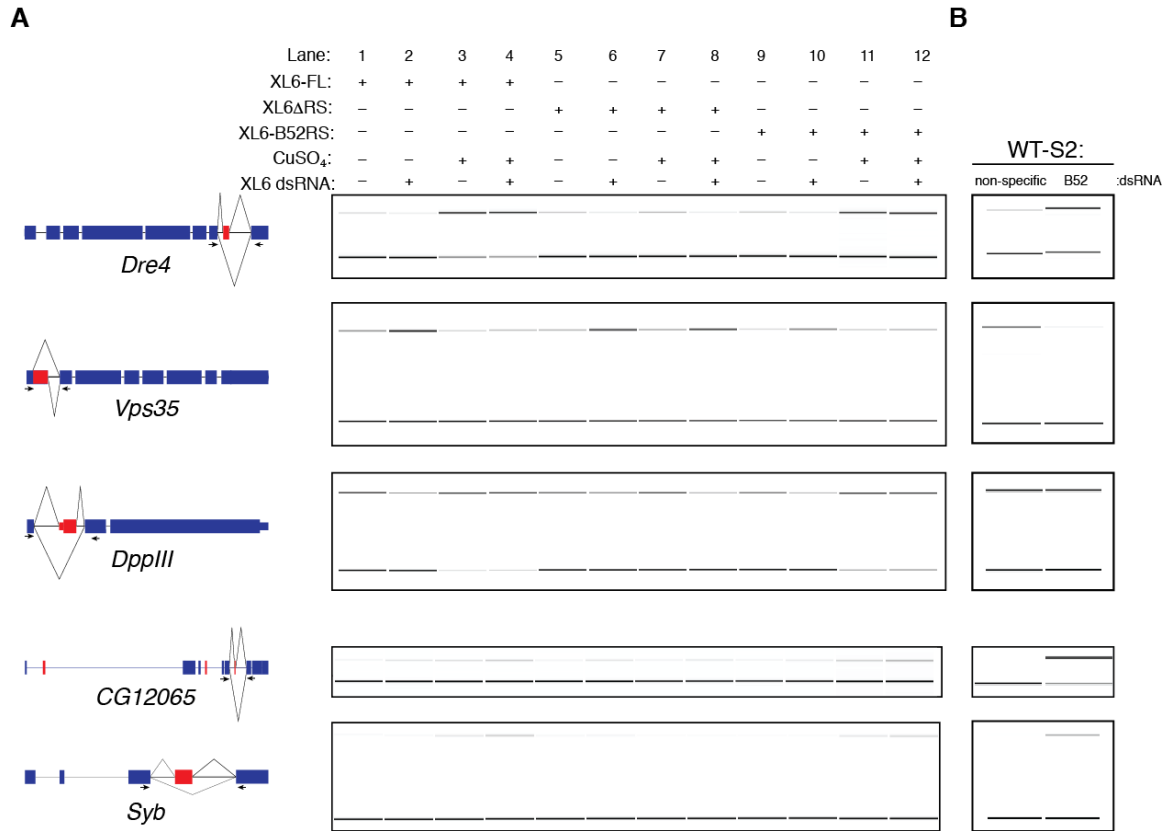


Figure 3.4: Specificity of SR protein AS regulation depends on RRM-domain, and requires the RS-domain. (A) Semi-quantitative RT-PCR of XL6 regulated targets (*Dre4*, *VPS35*, *DppIII*) and B52 regulated controls (*CG12065* and *Syb*) in S2 cells that are expressing wild-type XL6 or XL6 mutants. **(B)** Semi-quantitative RT-PCR of AS targets in the presence or absence of B52 RNAi in wild-type S2 cells.

However, over-expression of XL6 Δ RS fail to enhance the activation of the *Dre4* AS event, and AS isoform regulation is still sensitive to endogenous XL6 protein levels (Figure 3.4A; compare lanes 7 and 8 to lanes 5 and 6). Similarly, over-expression of XL6 Δ RS does not enhance the repression of the AS event in *Vps35*, and this event is also sensitive to endogenous XL6 protein levels. This demonstrates that, for *Dre4* and *Vps35*

AS events, over-expression of wild-type XL6 can functionally enhance AS, but this regulation requires an RS-domain.

When B52 protein level is reduced there is antagonistic regulation of *Dre4* and *Vps35* AS events. In addition to direction, it appears that when B52 is reduced there is a greater magnitude of splicing change. B52 regulates the largest number of AS events, and events regulated by B52 have an increased magnitude of splicing change as compared to XL6 depletion (Chapter two). The RS-domain of B52 is 41 amino acids larger than the RS-domain of XL6 (Figure 3.1). Previous studies have demonstrated that the splicing activity of bound SR proteins is directly proportional to the number of RS tetrapeptide sequences within the RS-domain, and this is related to the length of the RS-domain. In addition to a greater overall length of the RS-domain, B52 has 11 RS tetrapeptide repeats, while XL6 only has four. To test if the RS-domain from B52 can functionally substitute for the XL6 RS-domain we over-expressed a construct that has the B52 RS-domain substituted for the XL6 RS-domain. Over-expression of XL6-B52RS also enhances the activation and repression of *Dre4* and *Vps35* AS events, respectively. Similar to the over-expression of wild-type XL6, this enhancement persists regardless of endogenous XL6 protein levels (Figure 3.4A; compare lanes 11 and 12 to lanes 9 and 10). Interestingly, the direction of splicing change when XL6-B52RS is overexpressed mimics that of wild-type XL6, even though endogenous B52 affects *Dre4* and *Vps35* antagonistically of XL6. In addition, the magnitude of splicing change is less with the longer B52 RS-domain. This indicates that the SR protein RRM-domain, and not a specific RS-domain, dictates the direction of the splicing change. Furthermore, the magnitude of splicing change also depends on the binding location of the SR protein, and less with the length of the RS-domain.

The cassette exon in *DppIII* is activated in the presence of XL6, and a modulation of B52 protein level has no effect (Figure 3.4A; compare lanes 1 and 2). Over-expression of XL6-FL and XL6-B52RS, but not XL6 Δ RS, enhance the inclusion of the *DppIII* AS exon. Similar to the AS events in *Dre4* and *Vps35*, XL6-B52RS can enhance AS in the same direction as XL6-FL. However, AS events uniquely regulated by B52, and not XL6 (*CG12065* and *Syb*), there is little or no change in AS when XL6-FL, XL6 Δ RS, or XL6-B52RS is over-expressed (Figure 3.4A and B). This indicates that the specificity of targets is also determined by the RRM-domain, and not the RS-domain.

These results strongly suggest that the RS-domain of SR proteins is required for AS regulation, *in vivo*. Furthermore, RS-domains from different SR proteins can be functionally substituted, but do not dictate the direction of splicing change or the specificity of the targets. In addition, it seems that substituting a longer RS-domain with more RS tetrapeptide repeats from another SR protein does not correlate with the magnitude of splicing change. However, the location, or affinity, of binding to the target has the most influence on SR protein activity, and the SR protein RRM-domain defines these attributes.

3.4 Discussion

In this chapter, we analyzed the contribution of SR protein RS-domains to regulation of endogenous AS targets, *in vivo*. It is believed that the RRM-domain mediates binding to ESE motifs, while the RS-domain mediates protein-protein interactions with other splicing factors to regulate splice site selection. We demonstrate that the RS-domain of XL6 is required to control the regulation of the AS of the endogenous targets tested.

Previous, *in vitro*, studies in mammals suggested that the RS-domain of SF2/ASF is required for constitutive splicing, but is dispensable for concentration-dependent regulation of some AS [40]. Furthermore, it was shown that SF2/ASF RS-domain could be dispensed for regulation of some constitutive splicing targets [137]. While our study did not examine constitutively spliced targets, we examined targets that were sensitive to XL6 depletion, whereas the previous studies only examined targets whose splicing changes when dramatically high amounts of recombinant SR protein is added to the splicing reaction. This difference may allow our study to identify AS targets that are more dependent on the action of XL6, and therefore, more sensitive to XL6 RS-domain. Secondly, our study examined splicing changes of endogenous targets, *in vivo*, and the other studies examined splicing substrates *in vitro*. There are most likely are differences in the assembly of splicing complexes, and contribution of co-transcriptional splicing regulation, between the two different systems. Another *in vivo* study in DT40 B-cells demonstrated the SF2/ASF that had an RS-domain was essential for cell viability [138]. Reversible phosphorylation of the RS-domain has been demonstrated to affect SR protein RNA-binding specificity, and subsequent function in the splicing reaction. However, it has been shown that SR proteins lacking the RS-domain can still bind to the same sequence recognized by the phosphorylated full-length SR protein [93]. Therefore, it may be expected that over-expressing XL6 Δ RS may bind functional XL6 binding sites and have a dominant-negative effect. In our study we do not see any change in splicing patterns when XL6 Δ RS is over expressed. This could mean that XL6 Δ RS is not binding to the same binding site as XL6, or may not be able to bind RNA at all. Previous studies have demonstrated that phosphorylation of SR protein RS-domains are required for

splice-site recognition, and dephosphorylation is required for proper splicing catalysis [139]. It would be interesting to assay if XL6 Δ RS can still bind RNA, and if so, can it still bind in the same location, and with the same affinity, as wild-type XL6.

RS-domains from different SR proteins are shown to be functionally substituted for one another. Additionally, replacement of the RS-domain with RS dipeptides can substitute for the entire RS-domain [137]. RS-domains from different SR proteins, when fused to the bacteriophage MS2 coat protein RNA-binding domain targeted to the same region on a pre-mRNA, have differential strengths of splicing enhancement. The differences in strength were attributed to RS-domain length, and number of RS tetrapeptides [60]. In our study the RS-domain from B52, when substituted for the XL6 RS-domain, can functionally substitute in AS regulation of XL6 targets. This extends the observation that RS-domains can be functionally substituted for one another for regulation of AS targets, *in vivo*. The RS-domain of B52 is larger, and contains more RS tetrapeptide repeats. However, we do not observe an increase in magnitude of splicing change, it is actually lower, when the XL6-B52RS is over-expressed. Alternatively, the apparent reduced activity of XL6-B52RS could result from differences in protein folding, or conformation, in the SR protein fusion.

When examining targets that are regulated antagonistically by XL6 and B52, we show that the direction of splicing change is dependent on the XL6 RRM-domain. When the RS-domain from B52 is substituted for XL6 (XL6-B52RS), the direction of splicing change is in the direction of endogenous XL6, not B52, confirming that the RRM-domain of XL6 is responsible for the specificity of AS regulated targets. Conversely, XL6-

B52RS over-expression has little or no affect on AS targets regulated by B52 and not XL6.

The observations in this chapter demonstrate that SR proteins require an RS-domain to regulate AS, but RS-domains from other SR proteins can be functionally substituted, *in vivo*. Additionally, RNA-binding specificity is determined by the RRM-domain, and the binding location on the target transcript determines the pattern of AS regulation.

Therefore, it will be important to identify the binding location on target transcripts to fully understand the precise mechanism of SR protein AS regulation.

Chapter IV: Generation of Genome-wide RNA Binding Distribution For The SR Protein Family

4.1 Introduction

Throughout RNA metabolism, transcripts are associated with many RNA-binding proteins (RBPs) that coordinate multiple RNA processing and regulatory steps, which contribute to the ultimate expression and function of individual RNAs [140]. SR proteins play a significant role in multiple stages of an RNA lifecycle, and can regulate gene expression output at these different steps. The precise spatial arrangement of SR proteins on a target transcript are important for identifying direct targets, as well as uncovering how SR proteins regulate different RNA processes, including alternative pre-mRNA splicing.

Regulation of splicing is highly complex, and is the result of the combinatorial regulation of multiple splicing factors binding to several regulatory sequences in the RNA. For example, the SR proteins XL6 and Rbp1, in collaboration with Tra and Tra2, have been shown to bind to specific regulatory elements in the *Drosophila* transcript *doublesex* to regulate splicing [61]. Additionally, there are position-dependent effects on AS regulation. A mammalian SF2/ASF or SRp40 binding site was inserted at different locations on an alternative exon, and depending on the position, exerted different effects on the AS pattern [59]. These studies indicate that splicing can be regulated by multiple SR proteins, along with other RBPs, in a combinatorial fashion, and that the binding location of the SR protein in the target pre-mRNA determines whether it functions as an enhancer or silencer of splicing.

There have been several approaches taken to identify the physiological targets, and map protein-RNA interactions, to determine SR protein effects on the transcriptome.

Biochemical methods such as SELEX, electrophoretic mobility shift assay (EMSA), and RNA protection assays were first used to identify high-affinity binding sites for several SR proteins. These binding sites were purine-rich, and resembled ESE sequences [94]. Functional SELEX was developed to select for sequences that are bound, but will also promote splicing, using S100 extract. This strategy identified more redundant binding motifs between the SR proteins than the traditional SELEX method, indicating that binding specificity, *in vivo*, may depend on additional factors other than simple sequence recognition [58]. The motifs identified from traditional and functional SELEX have been used to scan for potential, endogenous, ESEs that are regulated by SR proteins.

One problem with these biochemical approaches is that they do not assay SR protein RNA-binding in a native cellular context, to endogenous targets. To preserve the cellular context, RNA immunoprecipitation (RIP) was combined with RT-PCR, or microarray analysis (RIP-CHIP), to identify nascent RNA-protein interactions. RIP-CHIP was used to identify the genome-wide binding locations of four members of the *Drosophila* hnRNP family of splicing regulators [74]. However, these methods are of low resolution, and can identify indirect RNA-binding interactions.

Most RBPs recognize short RNA motifs, and bind to several positions on a target RNA. Therefore, it is important to determine the landscape of RNA-binding of and RBP on a target RNA. To determine precise RBP binding sites, and reduce nonspecific binding, CLIP (UV cross-linking and immunoprecipitation) was developed to identify RNA targets of RBPs. CLIP involves the UV irradiation of living cells or tissue to induce the

formation of covalent cross-links only at sites of direct contact between proteins and RNA. Once the cells are lysed, protein-RNA complexes are immunopurified, followed by partial RNase digestion to generate short RNAs specifically bound to the RBP. These RNAs are subjected to reverse-transcription to generate bound cDNAs. In combination with high-throughput sequencing (CLIP-Seq), precise RBP binding locations, genome-wide, can be determined. CLIP-Seq was first used to uncover the binding landscape of the neural specific RBP, NOVA, in the mammalian brain [100]. It has also been used for other RBPs, including mammalian SF2/ASF and hnRNP family members [101, 112].

Although the resolution of CLIP-Seq is greater than RIP-CHIP, the resolution of binding site detection corresponds with the length of the digested RNA. In addition, it was demonstrated that the reverse-transcription reaction stops at many nucleotides that are cross-linked to the peptides that remain after proteinase K digestion, and are lost during the standard CLIP library preparation [105]. A modification of the CLIP protocol, termed iCLIP (individual nucleotide resolution CLIP), allows the truncated cDNAs to be sequenced. This identifies the location of the cross-link position, and the ability to obtain individual nucleotide resolution of cross-link sites. iCLIP replaces one of the RNA ligation steps in CLIP with a, more efficient, intramolecular cDNA circularization to capture the truncated cDNAs. iCLIP was first used to identify the binding footprint of hnRNP C to specific uridine rich tracts in order to regulate AS [104]. Subsequently, iCLIP has been used to identify the binding locations of two mammalian SR proteins (SRSF3 and SRSF4). This study found that these two SR proteins bind to unique locations on diverse types of RNAs, and can cross-regulate the expression of other SR proteins [107]. Although this study provides insights into SR protein function, it did not

examine SR protein binding positions, and their role in regulating AS. Additionally, the previous study only profiled the binding locations of two members of the SR protein family.

In this chapter, we determine the *in vivo* binding location, transcriptome-wide, for all eight members of the *Drosophila* SR protein family using iCLIP-Seq. Our analysis shows that SR proteins bind to an overlapping set of genes, but mostly bind to distinct locations on the target RNAs. We further demonstrate that SR proteins can bind to a functionally diverse set of RNAs, including intronless and non-coding RNAs (ncRNAs), and identify high-affinity binding sites of these targets. The combination of iCLIP-Seq and RNA-Seq reveal position-dependent regulation of AS, and interrelated RNA splicing maps for many SR proteins. These results characterize SR protein binding in respect to its role in AS regulation, but also implicate SR proteins in the regulation of other aspects of RNA processing, further supporting SR protein's widespread contribution to gene expression regulation.

4.2 Materials and Methods

4.2.1 Construction of 2x Flag-HA SR protein constructs

Full length SC35, SF2, SRp54, XL6, Rbp1, B52, Rsf1, and Rbp1-Like were PCR amplified and cloned into pDONR221 by BP reaction (Gateway cloning; Invitrogen).

Primers used for PCR for cloning into pDONR221:

SC35_FL_SP
5-GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGAGCAACGGTGGTGGT-3
SC35_FL_ASP

5-GGGGACCACTTTGTACAAGAAAGCTGGGTAGGAGCGACTGCGACT-3
 SF2_FL_SP
 5-GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGGGATCACGCAACGAG-3
 SF2_FL_ASP
 5-GGGGACCACTTTGTACAAGAAAGCTGGGTAATAGTTAGAACGTGA-3
 SRp54_FL_SP
 5-GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGGCTGGCGGCAACACC-3
 SRp54_FL_ASP
 5-GGGGACCACTTTGTACAAGAAAGCTGGGTAGGGCGAGTTGGAGAT-3
 XL6_FL_SP
 5- GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGTCGCGCCATCCGAGC-3
 XL6_FL_ASP
 5- GGGGACCACTTTGTACAAGAAAGCTGGGTAGTCCCTTGAAACGGA-3
 Rbp1_FL_SP
 5-GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGCCGCGATATAGGGAG-3
 Rbp1_FL_ASP
 5-GGGGACCACTTTGTACAAGAAAGCTGGGTAGAGTTGCAAGTTGAA-3
 B52_FL_SP
 5-GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGGTGGGATCTCGAGTG-3
 B52_FL_ASP
 5-GGGGACCACTTTGTACAAGAAAGCTGGGTAATCGTCCATGCTCTC-3
 Rsf1_FL_SP
 5-GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGTCCAGCATGGGTGAT-3
 Rsf1_FL_ASP
 5-GGGGACCACTTTGTACAAGAAAGCTGGGTAGAATCGATGGTTTCC-3
 Rbp1-Like_FL_SP
 5- GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGCCACGCTACCGTGAA-3
 Rbp1-Like_FL_ASP
 5-
 GGGGACCACTTTGTACAAGAAAGCTGGGTAATGGCGGTCCCTGGAGTCGGA-3

Next, the BP reaction was transformed into

DH5aext, the BP reaction was transformed into DH5GAGTCGGA-3Rbp1-Like were PC
 R amplified and cloned into pDONR221 by BP reactionet of RNAs, includinmetallothion
 ein promoter and a C-terminal 2x Flag and 2x HA tag (pMT-C2FL2HA) by LR reaction
 (Gateway cloning; Invitrogen). The LR reaction was transformed into
 DH5aetalllothionein promoter and a C-terminal 2x Flag and 2x HA tag (pMT-C2F

4.2.2 Generation stably transfected cell lines

400ng of pMT-C2FL2HA plasmid DNA containing either SC35, SF2, SRp54, XL6, Rbp1, B52, Rsf1, or Rbp1-Like, and 40ng of plasmid carrying the gene for blasticidin resistance (pCO-BLAST) were transfected into 2.0×10^6 *Drosophila* S2 cells, seeded in a 6-well tissue culture dish, using Effectene transfection reagent (Qiagen) according to the manufactures protocol. After 48hrs, blasticidin (Invivogen) was added to the transfected cells at 25µg/ml. Resistant cells were expanded, and maintained, in Schneiders media with 10%FBS and 25µg/ml Blasticidin.

4.2.3 iCLIP protocol

Performed iCLIP as previously described (Figure 4.1) [104]. 6.0×10^6 SR protein stably transfected S2 cells were seeded in a 10 cm dish, and protein expression induced by adding 100 µM CuSO₄ for 48hrs. Media was removed and cells were covered with ice-cold PBS, and subjected to UV irradiation (150mJ/cm² at 254nm). Cells were harvested into microtubes, and pelleted by centrifugation for 5 min at 1,000g. Cells were washed once with PBS, pelleted, and resuspended in 1ml lysis buffer (50mM Tris-HCl, pH 7.4; 100mM NaCl; 1% NP-40; 0.1% SDS; Protease inhibitory cocktail tablet (Roche)).

Prepare a 1/500 dilution of RNase I (Ambion, AM2295) Add 10µl of RNase I dilution and 2µl Turbo DNase, to the cell lysate. For control samples (-UV and +UV) make a 1/50 dilution of RNase I for a high RNase sample. Incubate the samples for exactly 3 min. at 37°C shaking at 1,100 rpm. Transfer to ice. Centrifuge at 13,000g for 20min at 4°C to clear the lysate.

For magnetic bead preparation, add 50µl of anti-FLAG M2 magnetic beads (Sigma, M882) per experiment to a fresh microtube. Wash beads 2x with lysis buffer. Add the

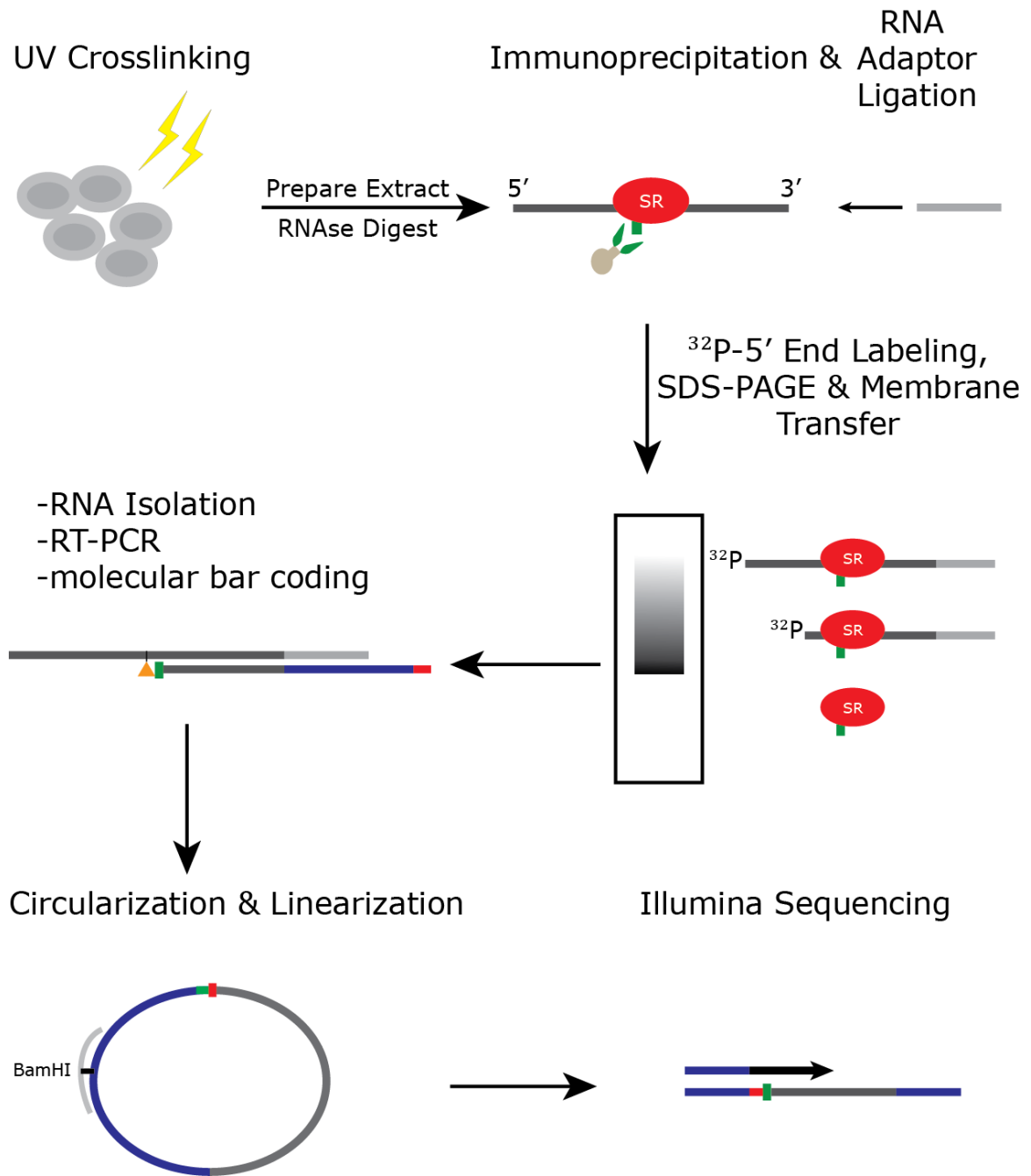


Figure 4.1: Schematic overview of iCLIP procedure. RNA adapter (light gray bar); adaptor sequence (blue bar); random molecular barcode (red bar); cross-linked polypeptide (orange triangle); cDNA region (Dark gray bar).

digested cell lysate to beads. Rotate the samples for 2hr at 4°C. Discard the supernatant and wash the beads 2x with 900µl high-salt buffer (50mM Tris-HCl, pH 7.4; 1M NaCl; 1mM EDTA; 1% NP-40; 0.1% SDS) and 2x with 900µl wash buffer (20mM Tris-HCl, pH 7.4; 10mM MgCl₂; 0.2% Tween-20).

For dephosphorylation of 3' ends, beads were resuspended in 20µl PNK mix (17µl water; 2.0µl PNK buffer [NEB]; 0.5µl PNK enzyme [NEB]; 0.5µl RNasin [Promega]) and incubate for 20 min. at 37°C. Add 500µl wash buffer and then wash with 1x high-salt wash buffer, followed by 2x with wash buffer.

For linker ligation to RNA 3' ends, beads were resuspended in 20µl ligation mix (2.0µl ligation buffer [NEB]; 2µl ATP [NEB]; 1.5µl 3' Linker [20µM]; 1.0µl RNA ligase [NEB]; 0.5µl RNasin [Promega]; 4.0µl PEG400 [Sigma]; 9.0µl water). Incubate at 16°C overnight. Add 500µl wash buffer and then was with 2x high-salt buffer and 2x wash buffer.

For labeling the 5' end, beads were resuspended in 10µl of hot PNK mix (7.5µl water; 1.0µl PNK buffer [NEB]; 1.0µl ³²P-g-n 10µl of hot PNK mix (7.5µl and incubate for 5 min. at 37°C. Hot PNK mix was removed and 22µl of 1x Nupage loading buffer (LDS; Invitrogen) was added to the beads. Beads were incubated for 10 min at 70°C. 2µl of the sample was saved for a western blot.

Samples were loaded on a 10% NuPAGE Bis-Tris gel (Invitrogen) according to the manufacturer's instructions. The gel was transferred to a nitrocellulose membrane using the iBLOT (Invitrogen) according to manufacturer's protocol. The membrane was rinsed

with PBS, wrapped in Saran wrap, and exposed to Kodak film MS from 30 min. to overnight.

The membrane was cut using the autoradiograph to isolate the protein-RNA complexes, and placed in a new microtube. 200µl of PK buffer (100mM Tris-HCl, pH 7.4; 50mM NaCl; 10mM EDTA) and 10µl proteinase K (Roche, 03115828001) was added to the membrane and incubated for 20 min at 37°C shaking at 1,100 rpm. Next 200µl PK buffer plus 7M Urea was added and incubated for an additional 20 min. at 37°C. The solution was collected and added to 400µl RNA phenol/chloroform (Ambion, 9722), and incubated for 5 min. at 30°C shaking. The sample was centrifuged for 5 min. at 13,000 rpm to separate the phases. The aqueous phase was transferred to a new tube containing 0.5µl glycoblue (Ambion, 9510), 40µl 3M sodium acetate pH 5.5, and 1ml 100% ethanol. RNA was precipitated overnight at -20°C. RNA was centrifuged for 20 min. at 15,000 rpm at 4°C, and the pellet washed with 500µl 80% ethanol and dried.

For the reverse-transcription reaction the pellet was resuspended in 10µl water, 1µl Rclip primer, and 1µl dNTPs (10mM). Reaction was incubated at 70°C for 5 min. then cooled to 25°C. Next, 8µl of RT mix (4µl FS buffer; 2µl 0.1M DTT; 1µl Superscript III [Invitrogen]; 1µl RNasin) was added and incubated 10 min at 42°C, 45 min. at 50°C, 15 min at 55°C, and 5 min at 90°C before cooling to 4°C. Volume was brought to 100µl with TE buffer and precipitated overnight as describe previously.

For size separation, cDNA pellet was resuspended in 6µl water and 6µl 2x TBE-urea loading buffer (Invitrogen), and heated at 80°C for 5 min before loading on a precast 6% TBE-urea gel (Invitrogen). The gel was run for 40 min at 180V as described by the

manufacturer. Three bands were cut at 120-200nt (high), 85-120nt (medium), and 70-85nt (low) and crushed in 400µl TE in a microtube. cDNA was allowed to diffuse for 2hr at 37°C. The supernatant was purified in a Costar SpinX column (Corning) containing two 1cm glass pre-filters (Whatman, 1823010) spun for 1 min at 13,000rpm. cDNA was precipitated overnight as described previously.

In order to circularize the cDNAs, the pellets were resuspended in 8µl ligation mix (6.5µl water; 0.8µl 10x CircLigase Buffer II; 0.4µl 50mM MnCl₂; 0.4µl CircLigase II [Epicentre]) and incubated for 1hr at 60°C. Next 30µl of oligo annealing mix is added (26µl water; 3µl FastDigest Buffer [Fermentas]; 1µl Cut Oligo [10µM]) and incubated 1 min at 95°C, then the temperature was decreased 1° every 20 seconds until 25°C is reached. Then 2µl BamHI [Fast Fermentas] was added and incubated 30 min. at 37°C. cDNA was precipitated as described previously.

To PCR amplify the library, the pellets are resuspended in 19µl of water and added to a PCR mix (1µl primer mix P5/P3 solexa [10µM each]; 20µl Accuprime Supermix [Invitrogen]). PCR was run using the following program: 95°C for 5 min, then 30-35 cycles of 95°C for 15 seconds, 65°C for 30 seconds, and 68°C for 30 seconds, then a final extension at 68°C for 3'. 1µl of PCR products were analyzed and quantified using a DNA 1000 chip on an Agilent bioanalyzer.

Primers:

L3 /5rApp/AGATCGGAAGAGCGGTTCAG/3ddC/

Cut_Oligo GTTCAGGATCCACGACGCTCTTCaaaa

P5/P3 Illumina Paired-end Primer mix

X33 = 5' Phosphate

Rclip1 X33AACCNNNNNAGATCGGAAGAGCGTCGTGgataCTGAACCGC

Rclip2 X33ACAANNNNNAGATCGGAAGAGCGTCGTGgataCTGAACCGC

Rclip3 X33ATTGNNNNNAGATCGGAAGAGCGTCGTGgataCTGAACCGC

Rclip4	X33AGGTNNNNNAGATCGGAAGAGCGTCGTGgataCTGAACCGC
Rclip5	X33CGCCNNNNNAGATCGGAAGAGCGTCGTGgataCTGAACCGC
Rclip6	X33CCGGNNNNNAGATCGGAAGAGCGTCGTGgataCTGAACCGC
Rclip7	X33CTAANNNNNAGATCGGAAGAGCGTCGTGgataCTGAACCGC
Rclip8	X33CATTNNNNNAGATCGGAAGAGCGTCGTGgataCTGAACCGC
Rclip9	X33GCCANNNNNAGATCGGAAGAGCGTCGTGgataCTGAACCGC
Rclip10	X33GACCNNNNNAGATCGGAAGAGCGTCGTGgataCTGAACCGC
Rclip11	X33GGTTNNNNNAGATCGGAAGAGCGTCGTGgataCTGAACCGC
Rclip12	X33GTGGNNNNNAGATCGGAAGAGCGTCGTGgataCTGAACCGC
Rclip13	X33TCCGNNNNNAGATCGGAAGAGCGTCGTGgataCTGAACCGC
Rclip14	X33TGCCNNNNNAGATCGGAAGAGCGTCGTGgataCTGAACCGC
Rclip15	X33TATTNNNNNAGATCGGAAGAGCGTCGTGgataCTGAACCGC
Rclip16	X33TTAANNNNNAGATCGGAAGAGCGTCGTGgataCTGAACCGC

4.2.4 High-throughput sequencing and mapping

High-throughput sequencing of iCLIP cDNA libraries from two replicate experiments for each SR protein were sequenced multiple times on an Illumina HiSeq 2000 with a 50nt run length. Sequence reads included a 5nt random barcode and a 4nt unique experiment identifier. The adapter sequences were trimmed using fastx_clipper. Next duplicate reads that had exact identity, including the experiment and random barcode, were removed using fastx_collapser. The resulting cDNAs (Referred to as CLIP Tags) were aligned to the *Drosophila melanogaster* genome (v5.22) using bowtie.

4.2.5 Identification of iCLIP-Seq clusters

Significant clusters were calculated by first determining the CLIP Tag number cutoffs using the Poisson distribution as identified in Polymenidou et al [141]. CLIP Tags were divided into Uxons (can be exonic) and Introns (only intronic). Nucleotides that had CLIP Tags above background ($p\text{-cutoff} \leq 0.01$) were identified as significant cross-linked nucleotides. Nucleotides that are within 15nt of one another were merged into a

CLIP cluster. Identified CLIP clusters from each replicate experiment were merged into experimental CLIP clusters. CLIP clusters genomic annotations were assigned based on gene annotations given by Flybase (v5.22).

4.2.6 Motif analysis

De novo motif finding was implemented using 1000 random SR protein CLIP clusters and the MEME-CHIP Suite that includes DREME. DREME is a motif discovery algorithm specifically designed to find short motifs from large datasets [142]. DREME was originally designed to identify transcription factor binding sites from CHIP-Seq data, but can be utilized to identify RBP binding motifs as well. The background used was the 1000 input sequences were shuffled preserving nucleotide composition. Consensus motifs that $E\text{-value} \leq 0.01$ were used. The E-value is the enrichment p-value times the number of motifs tested. P-values are calculated using the Fisher-exact Test for enrichment of the motif in the positive sequences.

MotifBS (<http://compbio.berkeley.edu/people/ed/motifBS.html>) was used to search the SR protein regulated AS genes with the identified motif, and return scores for the presence of the motif in the gene. The unaffected AS genes were also searched in order to obtain background scores. The cumulative distribution of scores for the regulated and background genes were plotted.

4.2.7 RNA splicing maps

iCLIP CLIP Tags were mapped to the affected cassette exon AS events detected by RNA-Seq (described in chapter two) at each nucleotide position in four regions (300nt

centered at each splice site). The relative density was calculated at each nucleotide in the regions, by dividing the number of CLIP Tags at a given nucleotide position by the total number of CLIP Tags in all regions. For each SR protein, cassette exons were separated into three groups, repressed, activated, and unchanged. Using R, we plotted at each nucleotide position the relative density of iCLIP CLIP Tags of activated, repressed, and unchanged cassette exons. The same was performed on competing acceptor (5' ss) splicing events.

4.3 Results

4.3.1 Mapping SR protein binding to RNAs at nucleotide resolution with iCLIP-Seq

Previous studies have demonstrated that, in order to control AS, members of the SR protein family bind *cis*-acting elements (ESEs and ISEs) in the target pre-mRNA.

However a detailed, global, description of bound RNAs for a majority of SR proteins has yet to be determined. Our previous study of AS changes when individual SR proteins are depleted, reveal that multiple SR protein family members co-regulate a majority of AS events (Chapter 2). One question raised by this observation is whether joint regulation requires the binding of multiple proteins, and if so, are the proteins binding to unique or overlapping regions on the target transcript. Therefore, it is important to identify the precise binding locations for all SR proteins on target transcripts to identify the mechanism of AS regulation by this family of conserved splicing regulators.

Furthermore, SR proteins have been implicated in multiple steps of RNA metabolism, so identifying global RNA targets, and the binding location on these targets, will provide insights into SR protein regulation of other RNA processing events.

We used the iCLIP method (Described in 4.1) to identify, direct, SR protein binding positions on RNAs, *in vivo*. All eight SR proteins were immunopurified via the FLAG tag encoded on stable transgenes transfected in *Drosophila* S2 cells. Expression of the SR protein transgene is under the control of an inducible metallothionein promoter. After 48hrs of induction, exogenous SR proteins are detected, and are not grossly overexpressed when compared to endogenous SR protein levels (Figure 4.2A). Previous analysis showed that the FLAG-tagged SR proteins could functionally rescue the AS patterns of the endogenous proteins (Chapter 3), and therefore should recapitulate interactions with target RNAs.

iCLIP experiments were performed, in duplicate, for all eight *Drosophila* SR proteins. Each SR protein was specifically and efficiently immunopurified from cell extracts, and SR protein-RNA complexes were isolated after *in vivo* UV-cross-linking. The purified SR protein-RNA complex was significantly reduced or absent when UV-cross-linking was omitted (Figure 4.2B). Cross-linked RNA was purified, reverse-transcribed, and PCR amplified in preparation for next-generation sequencing.

The resulting libraries were sequenced using Illumina Hi-Seq 2000, and generated between 168 and 432 million sequence reads per SR protein sample. Between 3.7 and 39.0 million sequences aligned uniquely to the *Drosophila melanogaster* genome, allowing for two mismatches, and were greater than 24nt in length. Next, removing identical reads, including the random and experimental barcode region, eliminated PCR amplification artifacts (Figure 4.3A). In addition, reads that had the same experimental and random barcode, were also collapsed in order to only count unique cDNA molecules (CLIP Tags). The resulting data set generated between 38,695 and 5.9 million unique

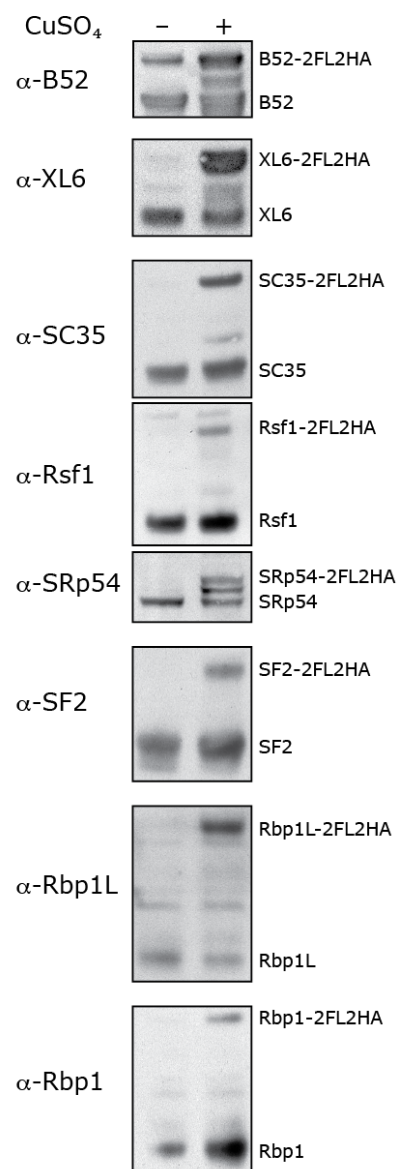
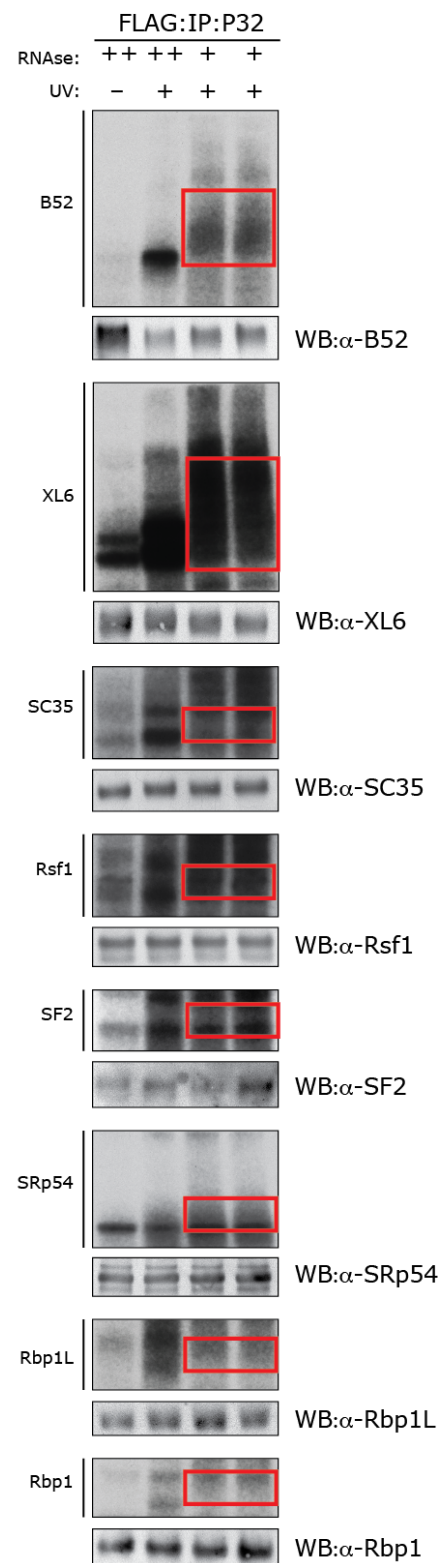
A**B**

Figure 4.2: Analysis of cross-linked SR protein-RNA complexes using iCLIP. (A) Western-blot of protein extracts after induction of tagged-SR protein for 48hr with 100 μ M CuSO₄. Extracts were blotted with antibodies that recognize the endogenous (Lower band) and exogenous (Upper band). **(B)** Autoradiograph of SR-RNA complexes digested with two concentrations of RNase, in the presence or absence of UV-cross-linking. The red boxes are the RNA that was excised and used for iCLIP-Seq library preparation. Below are western-blot of the immunopurified protein (WB).

CLIP Tags for each SR protein replicate (Figure 4.3B). We next used a cluster finding algorithm to identify significant SR protein binding sites from clusters of CLIP Tags, with a conservative threshold of $P \leq 0.01$. This stringent cut-off has the potential of removing true binding sites, but will limit our analysis to the most significantly bound regions. This analysis generated between 5,752 and 82,096 CLIP clusters for each SR protein replicate (Figure 4.3B). A limitation to iCLIP, and all CLIP methods, is that the libraries generated are of limited complexity. This is due to the low efficiency of UV-cross-linking, and the RNA ligation reaction. The iCLIP libraries generated for each replicate are most likely not saturated for all the bound RNA molecules, and only a few CLIP Tags represent many binding sites. However, all of the libraries were sequenced to saturation, indicating that the difference between replicates comes from the iCLIP procedure, and is not due to lack of sequencing. In order to minimize loss of true SR protein binding sites, we merged the identified CLIP clusters from each replicate. This produced between 16,004 and 102,04 CLIP clusters for each SR protein (Figure 4.3B). This analysis reveals that individual SR proteins bind a different number of RNA regions. XL6 binds the most while SF2 binds the fewest regions. This difference could be due to the iCLIP procedure (protein expression, cross-linking efficiency), but most likely is due to inherent differences in RNA binding distribution between different SR proteins.

A

Experiment	# of Reads	>24nt	Unique	% Aligned
B52	168,653,470	148,304,461	7,590,953	46.1
SF2	432,771,961	256,077,106	11,175,675	12.2
SRp54	383,981,671	256,645,564	19,932,676	31.5
XL6	230,405,486	211,043,353	31,585,884	47.5
Rbp1	389,532,969	323,752,140	31,703,936	25.8
Rbp1Like	432,765,937	374,140,215	39,014,689	39.5
Rsf1	278,067,059	147,805,072	6,103,062	35.8
SC35	301,847,745	150,638,277	3,774,665	36.2
B52-2/SC35-1	247,027,559	191,521,492	3,928,162	34.9

B

Experiment	Replicate	# of Molecules	Count of Multiplicity	Significant Clusters	Merged Clusters
B52	1	851,372	36,698,467	30,198	46,461
	2	158,425	73,871,729	19,279	
SF2	1	93,646	12,720,278	8,645	16,004
	2	87,567	18,831,057	8,213	
SRp54	1	357,013	47,272,991	16,067	38,425
	2	1,190,077	40,232,403	29,023	
XL6	1	3,610,312	47,898,509	64,040	102,040
	2	5,959,818	45,270,215	82,096	
Rbp1	1	602,232	50,257,100	23,722	38,563
	2	222,473	39,546,390	18,405	
Rbp1Like	1	1,280,424	79,938,741	43,258	64,539
	2	992,068	75,129,829	38,634	
Rsf1	1	197,788	32,690,359	24,069	43,406
	2	163,776	25,557,547	21,468	
SC35	1	119,212	30,940,202	13,523	18,704
	2	38,695	69,543,616	5,752	

Figure 4.3: iCLIP mapping information. (A) Number of RNA-Seq reads per experiment and alignment information. **(B)** Number of molecules were identified for each replicate using the random barcode. The number of significant CLIP clusters was calculated for each replicate, and then merged for each SR protein experiment.

4.3.2 Global landscape of RNA transcripts bound by SR proteins

Examination of SR protein CLIP clusters reveals that SR proteins bind to multiple locations across RNA sequences, and the same transcript can display cross-linking by multiple SR protein family members. This is exemplified when examining cross-linking of four SR proteins (B52, Rbp1, SRp54, and XL6) on two SR regulated transcripts. Both *Dre4* and *CG6084* display CLIP clusters for all four SR proteins (Figure 4.4A and B). While some clusters overlap, many CLIP clusters map to distinct regions of the transcripts. In addition, B52 and XL6 have many more CLIP clusters than Rbp1 and SRp54 on the *Dre4* transcript, but all four SR proteins have a similar number of CLIP clusters on *CG6084* (Figure 4.4A and B). These examples demonstrate that multiple SR proteins bind to target RNAs with different specificities. Although multiple SR proteins can bind overlapping regions, many regions show distinct binding for specific SR proteins.

To gain a better understanding of SR protein binding events, genome-wide, we profiled their binding on all annotated transcripts. We detected significant CLIP clusters in between 5,679 (SF2) and 9,952 (SRp54) genes for each SR protein (Figure 4.5A). We found clusters on greater than 90% of genes that are regulated at the level of AS by individual SR proteins. To test whether iCLIP captured binding on only the highly expressed genes, we compared the bound genes to our S2 cell expression data from RNA-Seq. Genes identified from RNA-Seq of S2 cells (Chapter 2) were separated into 4 quartiles of expression based on RPKM value. SR protein CLIP clusters were identified in genes at all four quartiles of expression, including genes not detected by RNA-Seq (Figure 4.5A). In general, most bound genes have one SR protein CLIP cluster, but on

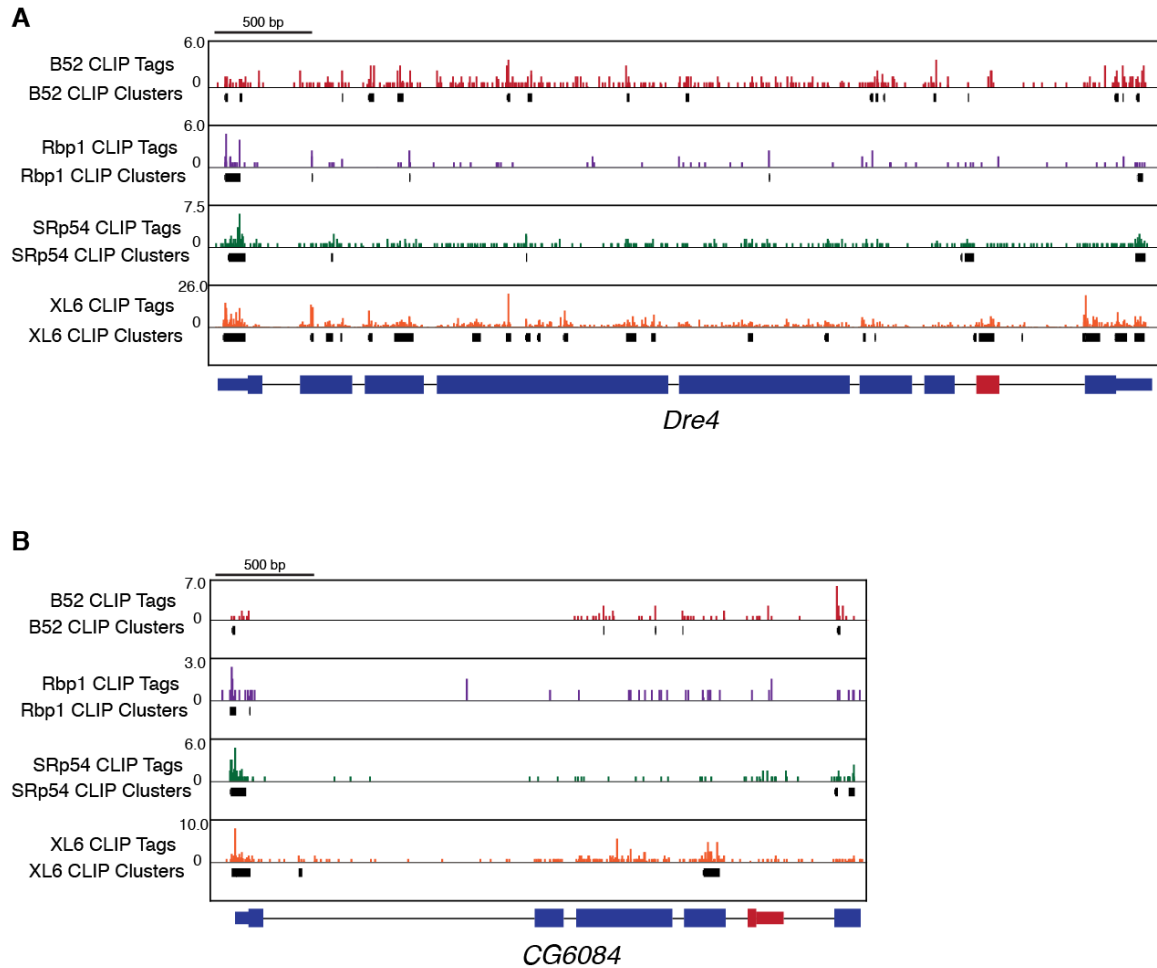


Figure 4.4: SR proteins bind to distinct positions in *Drosophila* RNAs. (A) *Dre4* gene with B52, Rbp1, SRp54, and XL6 CLIP Tags and CLIP Clusters. The numbers on the left represent the number of CLIP tags in the window. The alternatively spliced exon is in red. (B) *CG6084* gene with B52, Rbp1, SRp54, and XL6 CLIP Tags and CLIP Clusters. The numbers on the left represent the number of CLIP tags in the window. The alternatively spliced exon is in red.

average there are between 2.9 and 10.6 CLIP clusters per transcript depending on the protein, with a maximum of 60-313 CLIP clusters in *Ten-m* (SC35, XL6, Rbp1, B52, Rsf1), *Pum* (SF2, Rbp1L), and *Klar* (SRp54; Figure 4.5B and C). These three genes have extensive involvement in neurogenesis and contain large introns that need to be efficiently spliced. From this analysis, XL6 is a more promiscuous, and binds to more

RNAs, and to multiple locations, on the transcripts. Whereas, SF2 binds a smaller number of genes, and on average, fewer locations on the RNA.

A previous study using the two mammalian SR proteins SRSF3 and SRSF4 demonstrated that the two SR proteins bound to distinct genes, with little overlap of bound genes between the two [107]. Contrary to this study, we observed the binding of multiple SR proteins on *Dre4* and *CG6084* (Figure 4.4 A and B). Indeed, greater than 94% of the genes bound by each SR protein are also bound by at least one other SR protein family member (Figure 4.6A). However, when examining the overlap of the CLIP clusters, we observe a decrease in clusters that overlap on the bound targets (greater than 42% in all cases; Figure 4.6B). This suggests that while SR proteins bind to a similar set of transcripts, they bind to distinct, and specific, locations. One caveat is that since we are not saturating our cDNA libraries, this lack of overlap could be due to under sampling of bound RNA molecules. However, even the samples with the lowest amounts of CLIP clusters are bound to unique positions, and this observation of distinct binding is consistent with previous studies of SR protein binding to target RNAs.

Next, we examined the categories of transcripts, and the different functional regions within these transcripts, that are bound by SR proteins. Analysis of the frequency with which SR protein CLIP clusters were mapped to gene regions revealed that SR proteins bind to a high number of protein coding genes (> 98% CLIP clusters), within exonic sequences (> 53% CLIP clusters; Figure 4.7A). Despite the fact that introns make up a larger sequence space (62.5% of the genome), this enrichment of exonic binding reflects the known role for SR proteins as binding to ESE elements in pre-mRNA splicing, or other post-splicing activities. Within the exonic sequences, SR proteins mostly bind

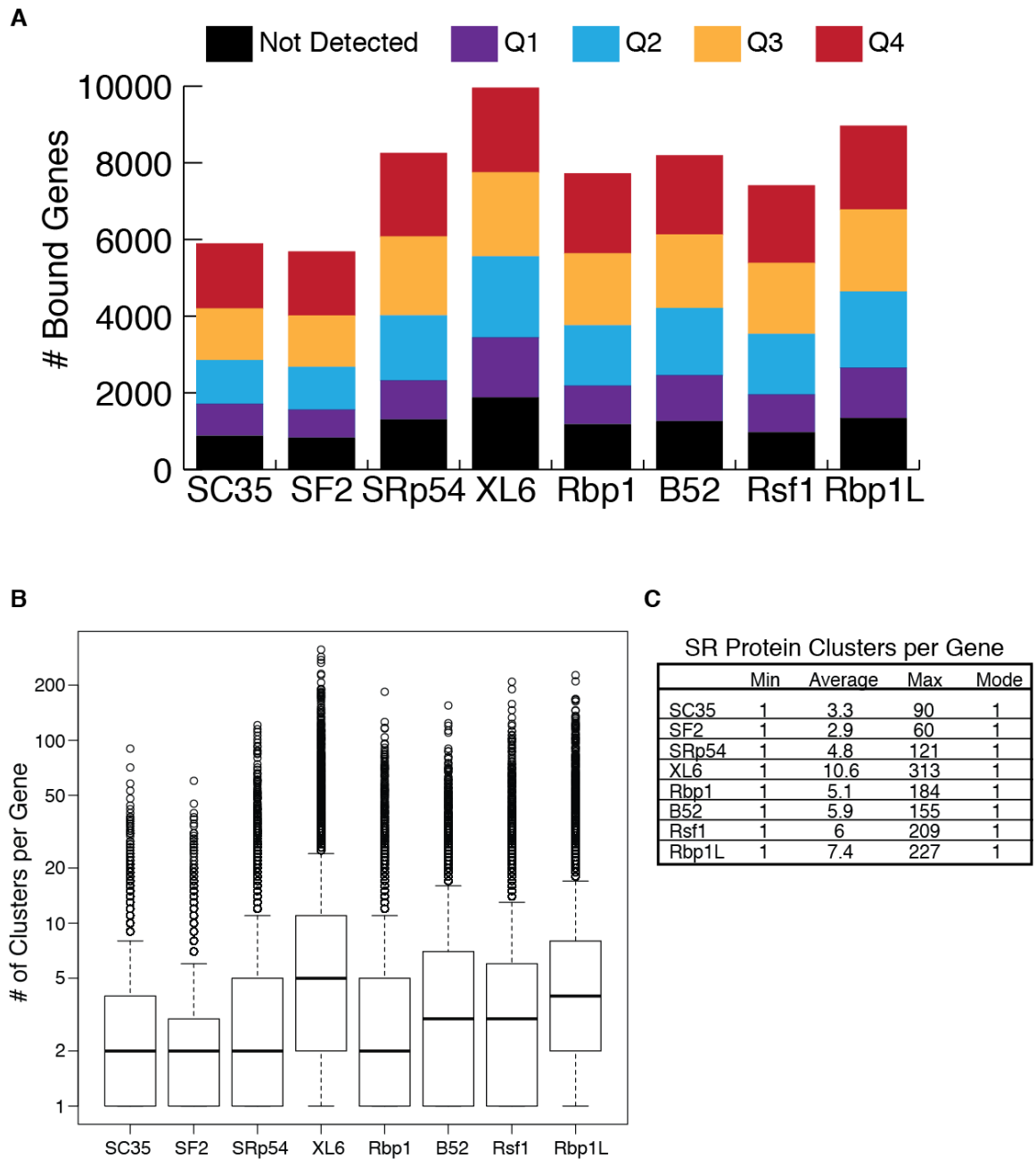


Figure 4.5: SR proteins bind a large number of genes . (A) Number of genes that contained an SR protein CLIP cluster, grouped by gene expression level. **(B)** Box plot illustrating the distribution of the number of CLIP clusters per gene for each SR protein sample. Black circles are the outliers. **(C)** Table displaying the minimum (min), average, maximum (max), and most frequent (mode) number of CLIP clusters per gene for each SR protein experiment.

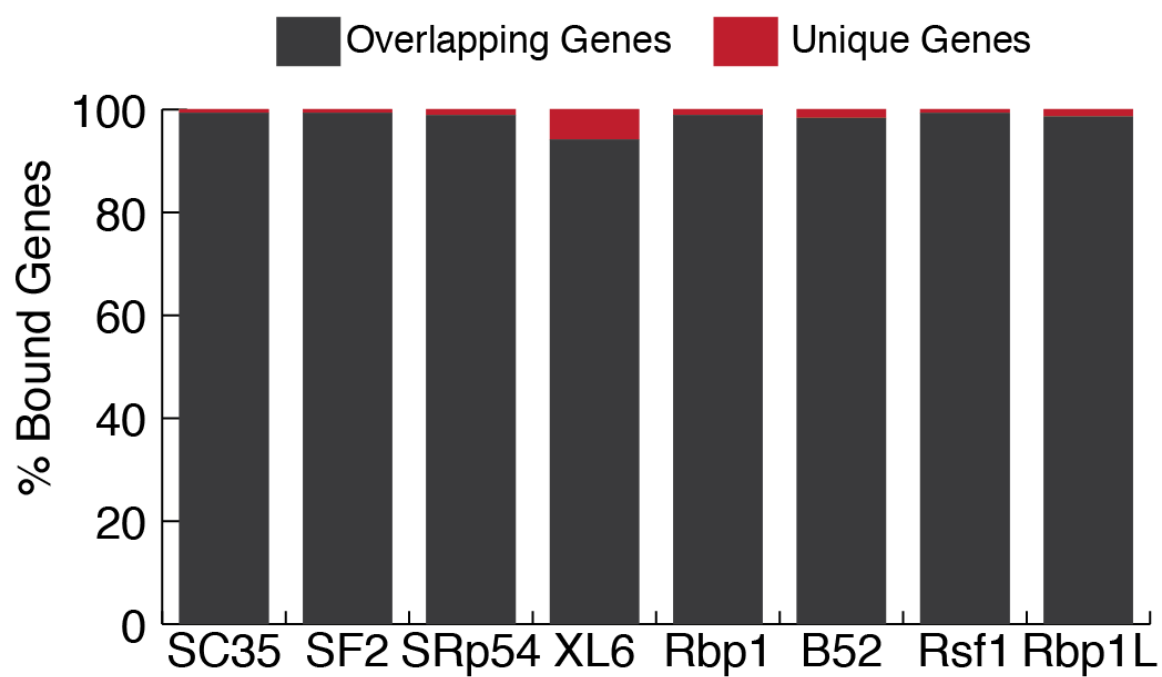
exons in the coding region, but have a slight enrichment for binding within the 5' UTR region (Figure 4.7A).

In addition to binding in protein coding genes, SR proteins can also bind non-coding RNAs (Figure 4.7A). SR protein binding to long ncRNAs for transport and stability, as well as, SR protein regulation of miRNA biogenesis has been previously shown [143, 144]. SR protein CLIP clusters were found in all the major types of ncRNAs (Figure 4.7B). The most abundant classes of ncRNAs with CLIP clusters were snoRNAs and tRNAs. snoRNAs are a class of small RNAs that guide RNA modifying enzymes, and tRNAs are a vital link between RNA and amino acids during protein translation. SR proteins have never been implicated in snoRNA or tRNA biogenesis. The extensive binding to these ncRNAs could indicate a novel role for SR proteins in the regulation of snoRNA and tRNA function, which could ultimately affect protein expression.

4.3.3 Consensus *in vivo* binding motifs for the SR proteins

Several SR protein-binding motifs have been identified using *in vitro* SELEX procedures. These binding sites are purine-rich sequences that closely resemble identified ESE elements, or 5' splice sites. However, using functional SELEX, it was revealed that SR proteins recognize more degenerate sequences, indicating that SR protein binding is more complex *in vivo*. In order to reconcile the differences in SR protein binding sites using different experimental methods, efforts to identify endogenous SR protein binding sites have been undertaken. Two studies have used CLIP methods to uncover the *in vivo* binding motifs of SF2/ASF (SRSF1), SRp20 (SRSF3) and SRp75 (SRSF4) in mammalian cell culture [107, 145]. To identify the, *in vivo*, binding motifs for the eight

A



B

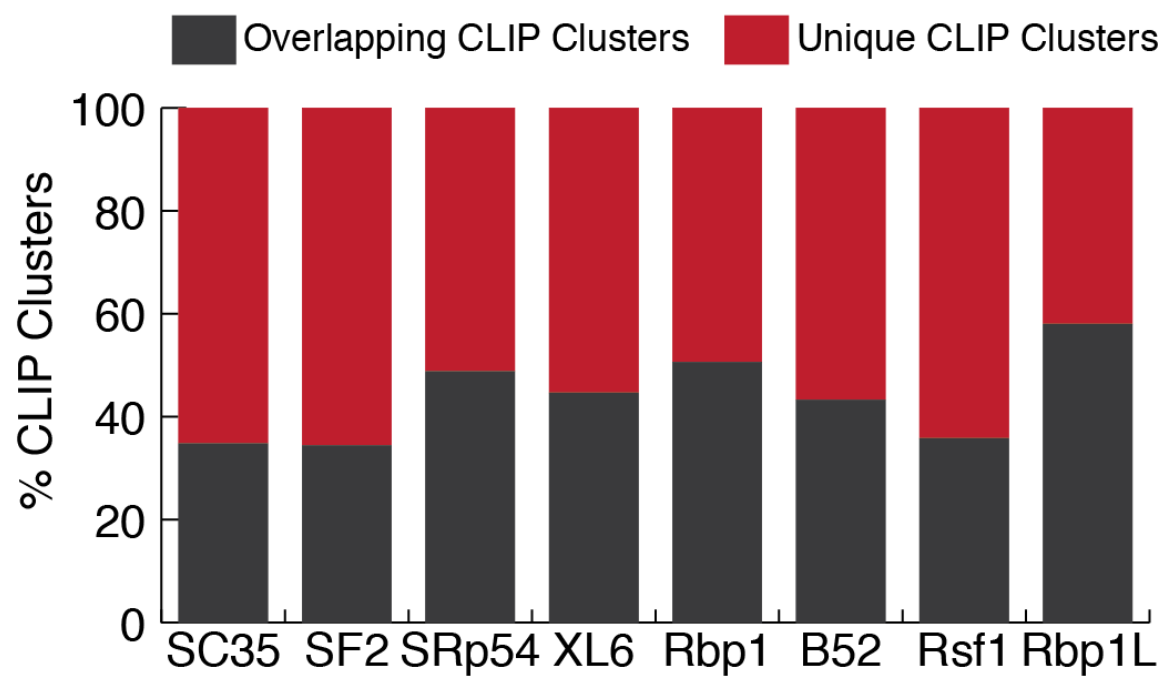


Figure 4.6: SR proteins bind a similar set of genes at unique positions. (A)

Percentage of bound genes that are bound uniquely (red) or by one or more other SR proteins (overlapping; gray). **(B)** Percentage of CLIP clusters that overlap with a CLIP cluster of another SR protein (overlapping CLIP clusters; gray) or do not overlap with another SR protein cluster (unique CLIP cluster; red).

Drosophila SR proteins, we used the MEME-CHIP suite, that allows for the identification of short binding motifs from large datasets, to identify enriched binding motifs from SR protein CLIP clusters (Figure 4.8). The binding motif for mammalian SRp20 was determined to be CU-rich excluding Gs, using, *in vivo*, binding data. One of the top five pentamers identified was UCAAC [107]. Mammalian SRSF3 has been shown to be very similar to XL6, Rbp1, and SRp54. Additionally, Rbp1 and XL6 have been demonstrated to be interchangeable for binding to the splicing enhancer in *Drosophila doublesex* [61]. Similar to their mammalian homolog, XL6, Rbp1, and SRp54 bind to similar CU-rich motifs that lack Gs (Figure 4.8).

Interestingly, Rbp1L, that has an RRM-domain that is 99% similar (differs by only two amino acids) to Rbp1, has a different binding motif than Rbp1 (Figure 4.8). Rbp1L's motif (GA[UA]G[AGU]) includes Gs, but is very similar to the ESE motif (GAAGAA) that was identified *in vitro*, and found to enhance splicing *in vivo* [89]. This reveals that even a slight difference in aa composition in the RRM-domain can influence binding site specificity.

Mammalian SF2/ASF was shown to bind to the purine-rich UG[GA][AU]G[AG]A motif *in vivo* [145]. Similarly, *Drosophila*, SF2 binds to a purine-rich G[GC]UGGA motif.

Binding motifs for the mammalian homologs of SC35, B52 and Rsf1 have not been

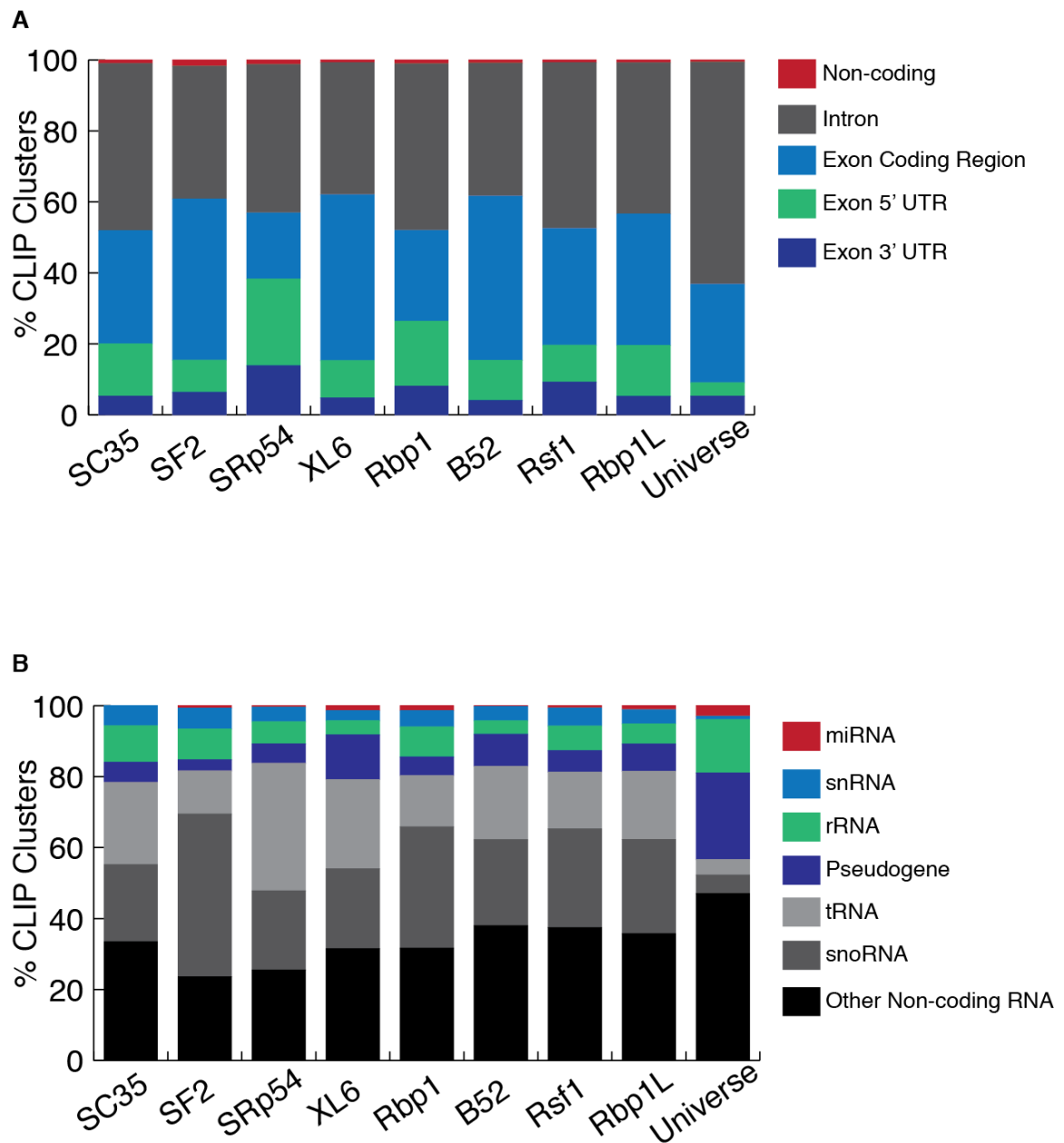
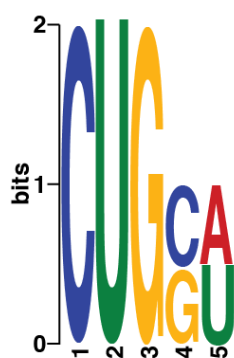


Figure 4.7: Distribution of SR protein CLIP clusters within RNA classes and transcript regions. (A) The proportion of CLIP clusters that mapped to different regions on bound transcripts for each SR protein. The proportion of genomic space is used as a reference (Universe). **(B)** The proportion of CLIP clusters that mapped to common types of non-coding RNA for each SR protein. The proportion of genomic space for each feature is used as a reference (Universe).

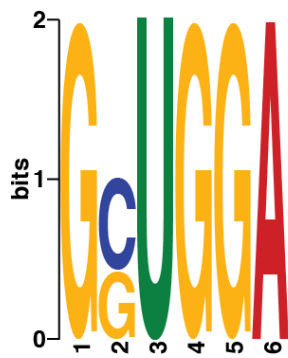
determined *in vivo*. However, *Drosophila* SC35 and B52 bind to a similar motif that was determined using SELEX [95, 146]. The binding motif for Rsf1 has never been determined, but is very similar to XL6 binding motif, and resembles an ESE (Figure 4.8).

Next, we examined the subset of genes that exhibit differential splicing when an individual SR protein is depleted (Chapter 2) for the extent of enrichment of the identified motif from iCLIP. We found that in all cases there was a high fraction of genes that had at least one occurrence of the SR protein-binding motif (Figure 4.9; blue line). We compared the distribution of this fraction, to the distribution of the fraction of genes that had a enriched motif score in the genes that are not AS regulated by the given SR protein, and found that as the motif score increased there was a smaller fraction of genes that contained the motif in the unregulated genes (Figure 4.9; red line). SC35, XL6, and B52 had only a slight enrichment over the background, while SRp54, SF2, Rbp1, Rsf1, and Rbp1L had a much larger enrichment (Figure 4.9). This indicates that there is a slight enrichment of identified SR protein consensus motifs in AS regulated genes as compared to genes not AS regulated. This observation supports the idea that there is a stricter requirement for a strong, or multiple, SR protein-binding motifs in AS regulated targets, compared to transcripts that are not regulated or constitutively spliced.

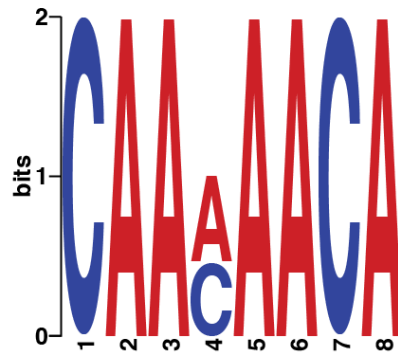
This analysis reveals that SR proteins bind to elements that resemble ESE motifs, and to motifs consistent with previously determined binding motifs for their mammalian homologs. In addition, Several SR proteins share a related binding motif (Rbp1 and SRp54), while others have a different, specific motif (SC35, SF2, XL6, B52, Rsf1, and Rbp1L). This suggests that SR proteins bind to specific motifs in order to regulate AS, and other RNA processing events.



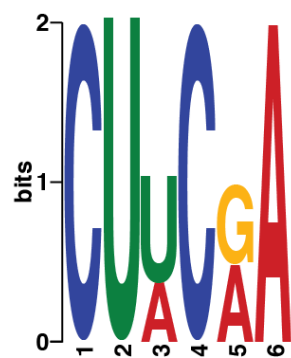
SC35



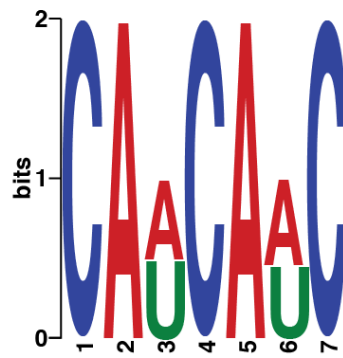
SF2



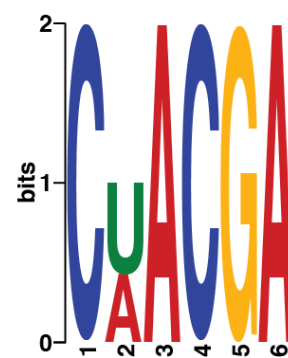
SRp54



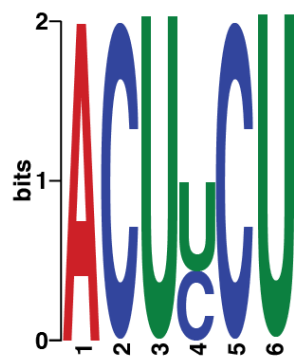
XL6



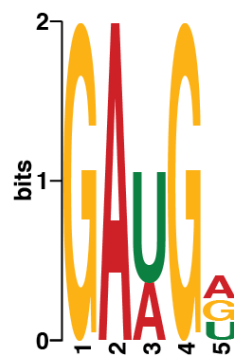
Rbp1



B52



Rsf1



Rbp1L

Figure 4.8: SR proteins bind to specific motifs, *in vivo*. Graphical representation of consensus binding motifs derived for each SR protein. Height of each bar shows the information content at each position of the binding motif in bits (log-odds in base 2).

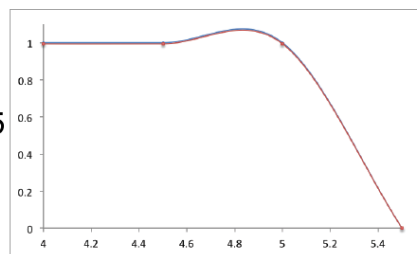
4.3.4 SR protein RNA splicing maps

In order to regulate pre-mRNA splicing, previous studies suggest that SR proteins bind to pre-mRNAs, primarily within exons, to recruit core spliceosomal components to nearby splice sites. Our previous analysis of CLIP clusters revealed a bias towards binding in exonic regions of annotated transcripts (Figure 4.7A). To map SR protein-binding positions, within exons and introns, we mapped the relative density of SR protein CLIP Tags 300nts around all 3' and 5' splice sites (Figure 4.10). As expected, mapping of SR protein CLIP Tags showed a very strong preference for exonic regions for all SR proteins. In addition, the density of CLIP Tags was more pronounced in exonic regions within 75nts of the acceptor or donor splice sites. SRp54 had prominent peaks within the intronic regions 10-20nt upstream of the 3'ss suggesting that SRp54 interaction with intronic sequences could regulate splicing. In conclusion, SR proteins preferentially contact exonic sequences, with an increase in density close to the splice sites. This is consistent with the model that SR proteins regulate pre-mRNA splicing by functioning as exonic splicing enhancers.

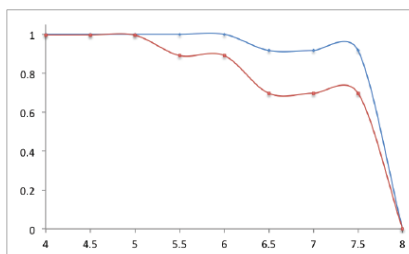
To determine the pattern of position-dependent binding of the SR proteins in relation to their role in regulating AS, we mapped the relative density of CLIP Tags on SR protein activated or SR protein repressed AS events identified for each SR protein (Chapter 2).

Fraction of Genes

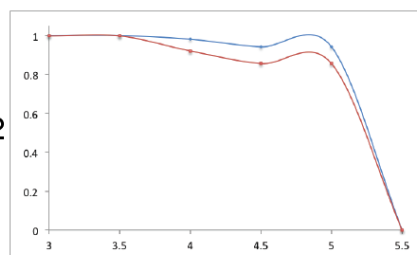
SC35



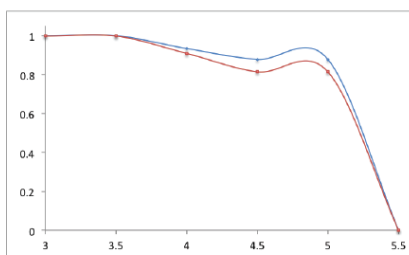
Rbp1



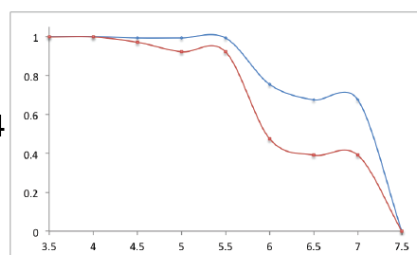
SF2



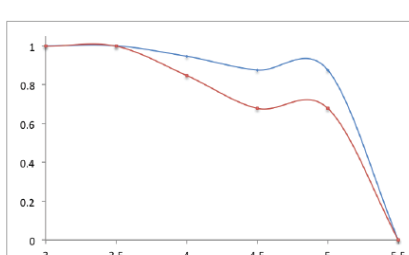
B52



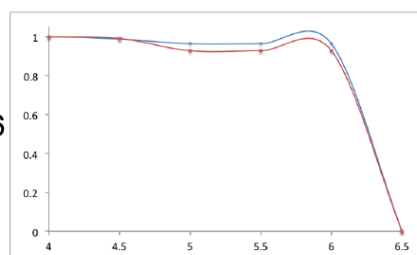
SRp54



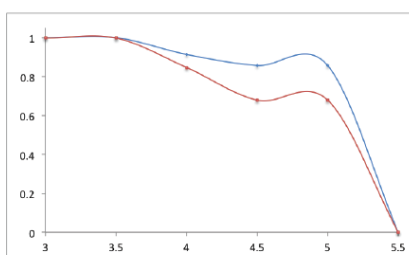
Rsf1



XL6



Rbp1L



Motif Score

— Regulated

— Not Regulated

Figure 4.9: Distribution of SR protein motifs in regulated AS targets. Cumulative distribution of the presence of at least one occurrence of the SR protein consensus motif in the genes identified as being regulated by individual SR proteins at the level of AS from RNA-Seq (Chapter 2; blue bar) and not regulated (red bar).

To calculate the density we divide the number of CLIP Tags at any given nucleotide (400nt around each splice site) by the total number of CLIP Tags across the alternative region. The average relative density is then calculated for all AS events at every position. A similar density was also calculated on ASEs that are not affected when an SR protein is depleted. First, we examined the binding position on regulated cassette exons for the four SR proteins that had the largest number of bound, and regulated, cassette exon events (Figure 4.11). This analysis reveals that all four SR proteins have higher binding density on activated cassette exons than repressed exons. In contrast, there is a higher density of binding to intronic and flanking constitutive exons when the SR protein is involved in repression of the alternative exon. Similar to our previous analysis, SR protein binding is clustered near the splice sites. Although, these common patterns of binding exists, individual SR proteins still have unique patterns of binding along the AS transcript. For example, SRp54 has more broad binding on the alternative exon, whereas B52, XL6, and SF2 have more specific, defined binding locations (Figure 4.11).

These observations were extended when we examined the SR protein binding distribution on competing donor AS events (Figure 4.12). SRp54, XL6, and SF2 displayed a higher density of binding near the distal 5'ss when the SR proteins activate its usage. Whereas, there is higher binding density near the proximal 5'ss when the three SR proteins activate its usage. B52, SRp54, and SF2 also had high binding density near the 3'ss when they

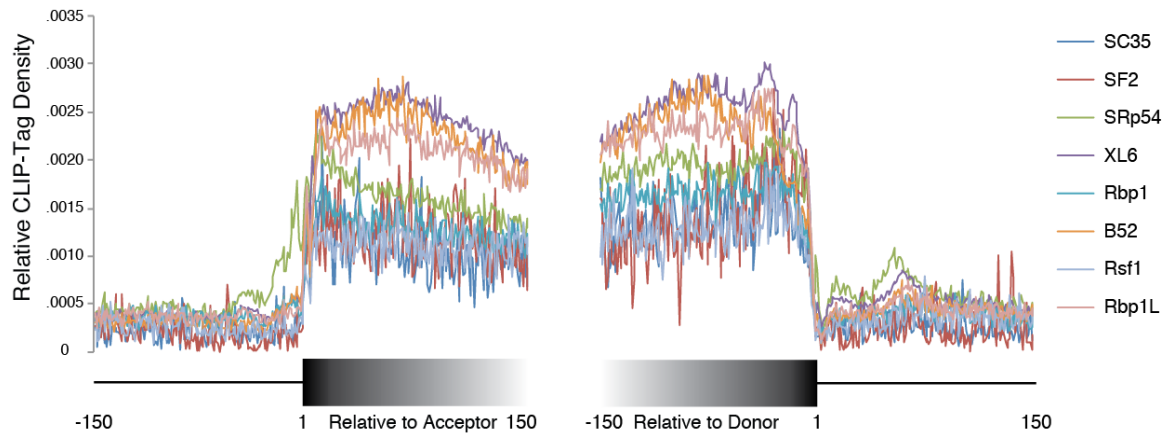


Figure 4.10: SR protein binding near splice sites. Relative density of CLIP tags 150nt upstream and downstream of the splicing acceptor (3'ss) and splicing donor (5'ss) for each SR protein experiment. Wide gray bar is exon sequence and thin gray bar is intronic sequence.

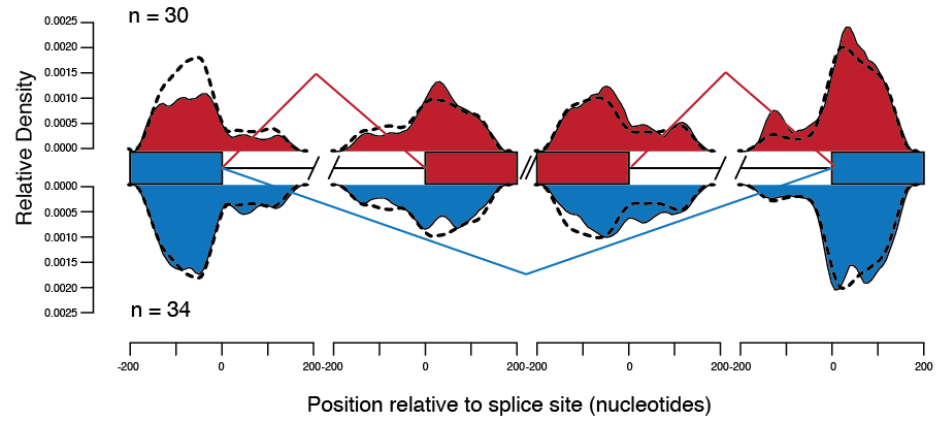
activated usage of the distal 5'ss (Figure 4.12). This could indicate that 3'ss selection can have an influence on 5'ss selection.

This analysis reveals that SR proteins bind to specific locations on their target transcripts to regulate AS. We observed that SR proteins are bound near the splice sites they activate, consistent with the model that SR proteins bind to ESE elements to recruit core components of the spliceosome to the proximal splice sites for splice site activation. SR proteins are bound to multiple locations on their targets, not ruling out the possibility that interactions with themselves, or other splicing regulators, are vital to their regulation of AS.

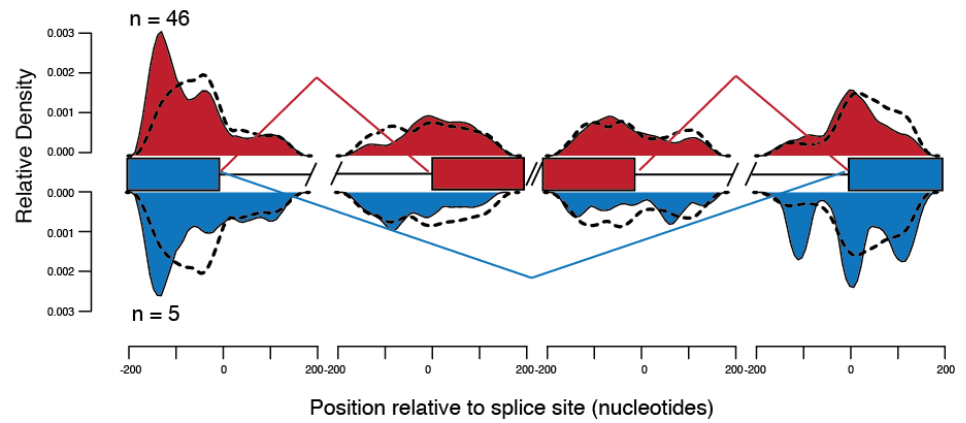
Unlike RNA-maps determined for NOVA and FOX splicing regulators, the RNA-maps for the SR proteins are more widely distributed along the target transcripts [100, 113].

This could reflect SR protein control of constitutive and alternative splicing of

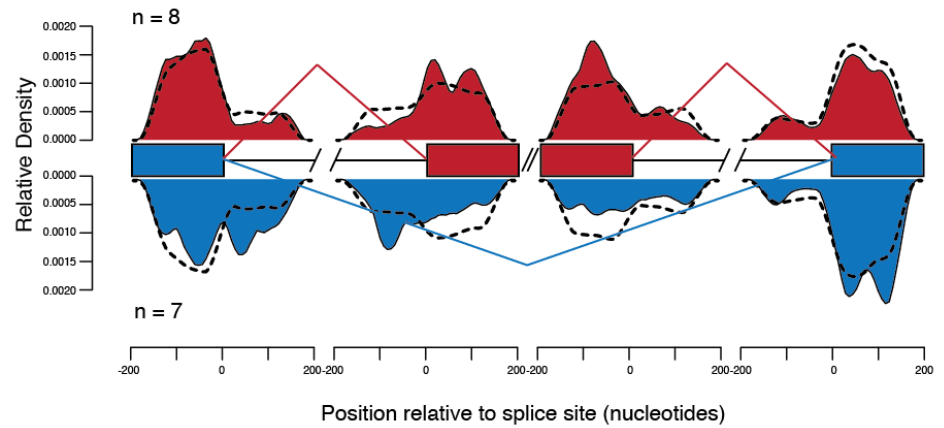
B52



SRp54



XL6



SF2

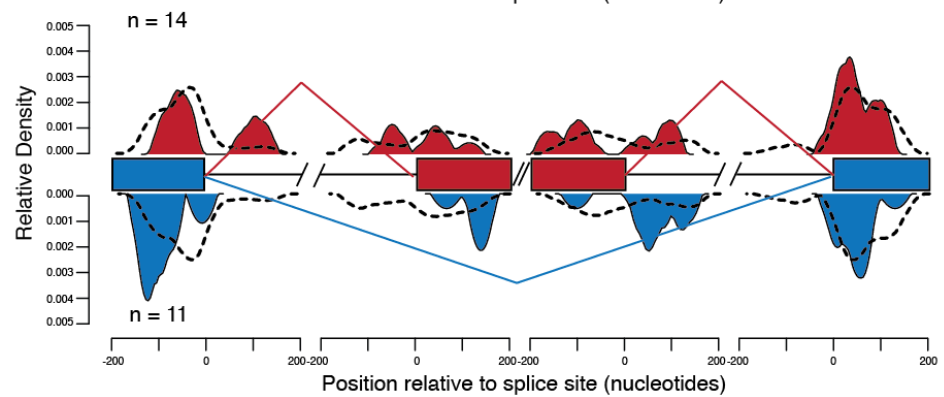


Figure 4.11: SR proteins bind in a position-specific manner around cassette exons. RNA splicing maps showing the relative density of CLIP tags proximal to activated (positive y-axis, red), repressed (negative y-axis, blue), and unaffected cassette exons (dashed line). The x-axis represents the nucleotide position relative to each splice site. The alternative exon is red and the flanking constitutive regions are shaded in blue. The total number of regulated events used for the analysis are listed (n).

many transcripts, as well as the more general control of other steps of RNA metabolism.

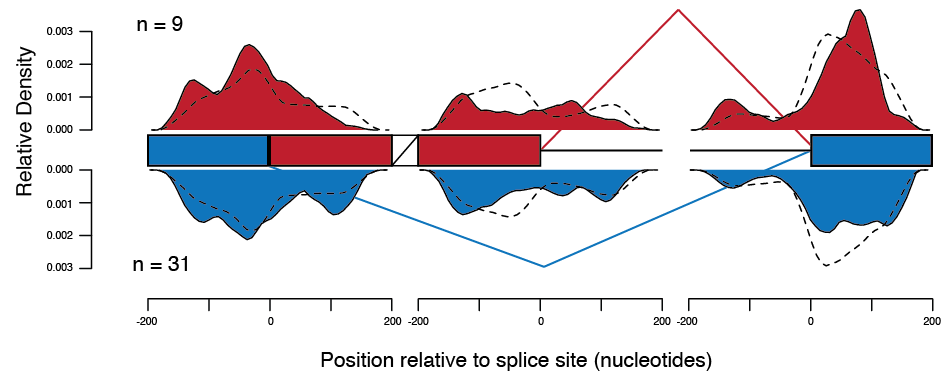
SR proteins are dynamically associated with many different complexes throughout RNA's lifecycle, and may bind to different locations to carry out multiple RNA processing activities, independent of AS. NOVA and FOX splicing regulators are tissue-specific splicing regulators; therefore their RNA-maps are a reflection of their specific function on a limited set of targets. A similar observation of broad binding distribution was made when RNA-maps were generated for the ubiquitously expressed hnRNP proteins in mammals [112].

4.4 Discussion

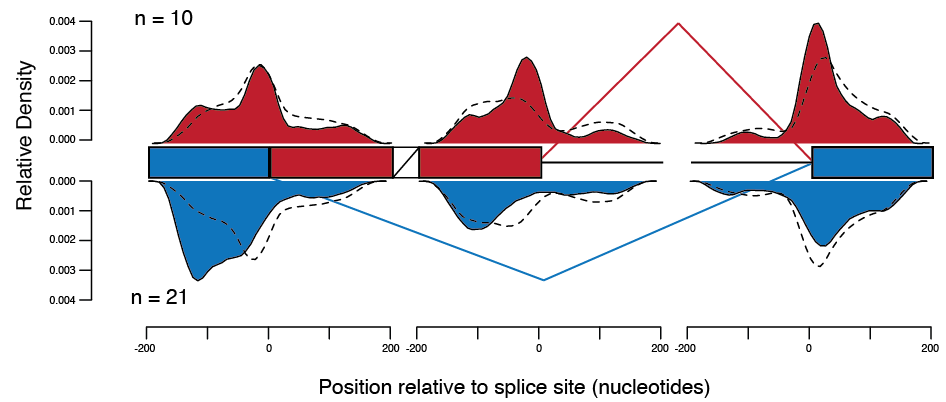
Recent studies using CLIP techniques to profile direct protein-RNA interactions have been instrumental in contributing to the understanding of post-transcriptional regulation. CLIP coupled with next-generation sequencing has allowed for the determination of, *in vivo*, protein-RNA interactions genome-wide.

CLIP-Seq and iCLIP-Seq have been used to profile the protein-RNA interactions of multiple splicing regulators, including three members of the mammalian SR protein family, but analysis of SR protein binding, and its role in regulating AS, has never been studied globally [101, 107, 112]. Using iCLIP-Seq, it was determined that SRSF3 and

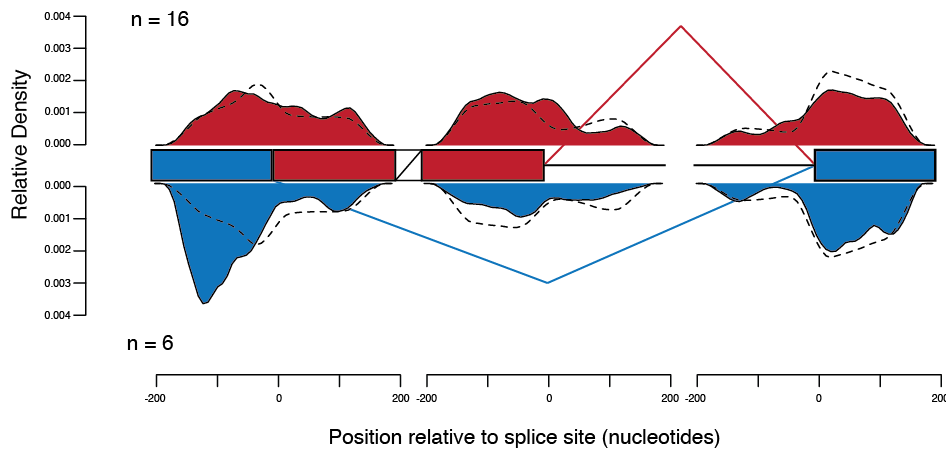
B52



SRp54



XL6



SF2

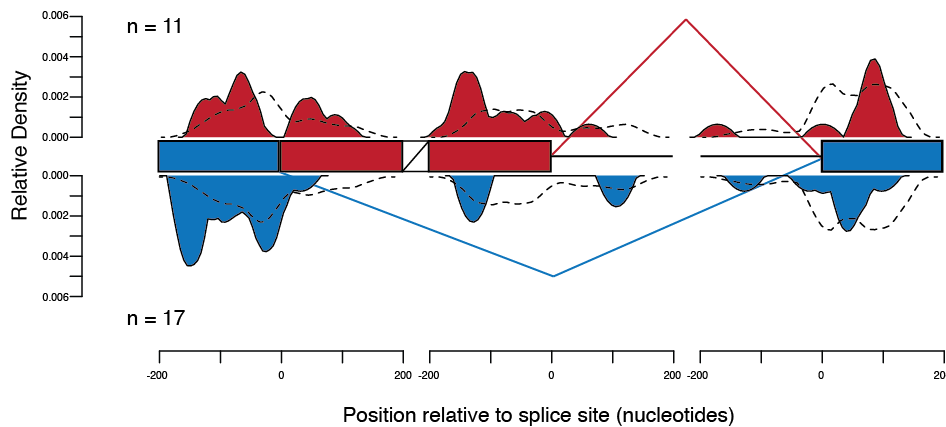


Figure 4.12: SR proteins bind in a position-specific manner around competing 5' splice sites. RNA splicing maps showing the relative density of CLIP tags proximal to activated distal 5'ss (positive y-axis, red), activated proximal 5'ss (negative y-axis, blue), and unaffected competing 5' splice sites (dashed line). The x-axis represents the nucleotide position relative to each splice site. The alternatively include region is red and the flanking constitutive regions are shaded in blue. The total number of regulated events used for the analysis are listed (n).

SRSF4, two mammalian SR proteins, bind to distinct, non-overlapping, sets of transcripts. In addition, they were determined to have very different RNA-binding motifs. One weakness of this study is that it did not profile RNA-binding with regard to the SR protein's role in regulating AS. Furthermore, only two members of the mammalian SR protein family were profiled, so it is not clear if observations for these two SR proteins can be extended to the entire family. In order to gain insight into how SR proteins regulate AS comprehensively, we used iCLIP-Seq to profile the RNA-binding landscape of all eight identified *Drosophila* SR proteins in S2 cells.

We identified the nucleotide-resolution mapping of, *in vivo*, SR cross-link sites. Using a statistical algorithm to identify significant clusters of enriched SR protein binding, we found that SR proteins bind to, a large set, of similar transcripts. In addition, we found that while different SR proteins can redundantly bind to the same location, most SR proteins interact with different RNA sequences. This finding is consistent with the observation that a majority of AS events are regulated, in a combinatorial fashion, by multiple members of the SR protein family (Chapter 2).

Global profiling of the SR protein-binding location on the target genes reveals that SR proteins bind, largely, to exonic sequences, and have a slight bias for binding in the 5'

UTR. This observation is consistent with previous analysis that SR proteins bind to enhancer elements in exonic regions, and act as general activators of splicing.

Interestingly, we observed SR protein binding within non-coding RNAs. Many of these non-coding RNAs are not processed by the spliceosome, but are regulated by other processing pathways. Previous studies have demonstrated that SR proteins can regulate additional RNA-processing events, such as miRNA biogenesis [144]. iCLIP-Seq revealed that SRSF3 and SRSF4 interact with long ncRNAs, as well as, short ncRNAs, such as snoRNAs. Indeed, one of the major over-represented groups of ncRNAs bound in our study were the snoRNAs. SR proteins are likely required for the splicing of all introns, including snoRNA-containing introns. Even though interaction between SR proteins and snoRNA has previously been described, the function of SR proteins, and the processing of snoRNAs are unclear. An interesting avenue for further studies would be the investigation of the activity of SR proteins in the release of the snoRNA from the intron lariat, or, the role of snoRNAs in co-regulation of AS.

The iCLIP data provided a large number of direct SR protein binding sites, enabling us to derive, *in vivo*, RNA-binding motifs. Therefore, we identified consensus-binding motifs for each SR protein. We found that, *in vivo*, SR proteins bind purine-rich sequences that resemble ESE sequence elements. While some SR proteins share a related motif, others have distinct RNA-binding specificities. The RNA-binding motifs for many of the mammalian SR protein family members have been identified using, *in vitro*, SELEX procedures. The motifs generated often vary, based on the experimental protocol used. Also, because the SELEX sequences are so short, it is difficult to use the identified motif to find *in vivo* targets. The sequence motifs identified in this study are very similar to the

binding sequences derived for the SR protein mammalian homologs using SELEX and CLIP techniques. This indicates that SELEX is a useful tool for identifying binding specificity alone, and that the SR protein binding motifs are conserved between *Drosophila* and mammals. The specificity and accuracy of the derived binding motifs has yet to be tested further. It would be interesting to do mutational studies, to determine how strict a requirement each derived motif is for SR protein binding and function.

Finally, we analyzed SR protein binding in respect to its role in regulating pre-mRNA splicing. Although SR protein binding is not limited to exons, we found more abundant SR protein binding within exons, with increasing density proximal to the splice sites. These data agree with model that SR proteins bind primarily to ESEs, and promote adjacent splice site recognition.

When comparing the SR protein binding locations to the outcome of AS, we found that SR proteins regulate AS in a position-dependent manner. In particular, cassette exons that are activated by SR proteins have a higher frequency of binding, and cassette exons repressed by SR proteins have a low frequency of SR protein binding, with higher binding density in the flanking introns and constitutive exons. This observation of SR protein binding near activated splice sites is extended when we analyzed SR regulated competing donor events. There was higher density of binding near the distal 5'ss when the SR proteins were responsible for its activation. These analyses provide insights into the mechanism of SR protein AS regulation.

Our global analysis of SR protein-RNA binding in this chapter shows that SR proteins interact with many of the same transcripts, but bind at different locations to regulate post-

transcriptional processing. SR proteins have specific, conserved, RNA-binding motifs, and display position-dependent binding in regards to AS regulation. Although primarily known as splicing regulators, the observation of binding to diverse classes of RNAs, including ncRNAs, implies SR proteins may have a broad, yet unidentified, role in many RNA processing events.

Chapter V: Perspectives and Conclusions

5.1 SR proteins and human disease

The precise control of AS, and other RNA processing steps, by RBPs is essential for proper protein expression in higher eukaryotes, and is utilized as a post-transcriptional mechanism of gene expression regulation. Computational predictions and transcriptome profiling of many normal and diseased tissues indicate that a high fraction of diseases are caused by mutations that alter splicing. It has been estimated that 15% of mutations that cause genetic disease affect conserved splicing signals, including elements recognized by SR proteins [17].

Elevated SR protein levels have been observed in a variety of human tumors, and correlate with cancer progression [25]. SRSF1 has been shown to regulate Ron (macrophage-stimulating 1 receptor) splicing that produces a constitutively active isoform, which contributes to increased cell motility and tumor metastasis [147]. HIV-1 uses, multiple, competing 5' and 3' splice sites to generate more than 40 mRNAs. Various SR proteins have been shown to regulate these splicing events in HIV-1, and may be a target for inhibition [148]. The inherited neurodegenerative disorder SMA results from a deletion of the SMN1 locus. A single polymorphism in the paralogous SMN2 gene increases skipping of exon 7, resulting in a decrease in the production of functional SMN protein. The single nucleotide change is thought to disrupt an SRSF1 ESE, which prevents SR protein binding, and subsequent activation of inclusion of exon 7 [18]. SR proteins have also been demonstrated to be autoantigens in systemic lupus

erythematosus (SLE), and have regulatory roles in the splicing of CFTR, that is the gene mutated in cystic fibrosis patients [149, 150].

Disruption of SR protein function contributes to the progression of many human diseases, however the complexity of the RNA targets, specificity of RNA-binding, and precise mechanism of AS regulation remains poorly understood for many members of the SR protein family. With the emergence of genome-wide methods to profile changes in AS, and detecting protein-RNA interactions, a global understanding of SR protein function is possible. In this dissertation we present genome-wide analysis of AS regulation, and RNA-binding information, for the full family of SR proteins. Hopefully, a clear understanding of SR protein regulation of AS can contribute to the development of novel therapies, or diagnostic approaches, for human disease.

5.2 Combinatorial regulation of AS by the SR proteins

SR proteins are identified by having similar protein domain architecture, and their ability to complement splicing-deficient S100 extracts. Since multiple SR proteins could substitute for one another in the splicing of *in vitro* substrates, it was thought that SR proteins were redundant regulators of splicing [130]. However, the observation of embryonic lethality for many individual members of the SR protein family, and further studies demonstrating that specific SR proteins were required for the regulation of splicing of certain pre-mRNAs, a specific role for individual SR proteins in splicing was uncovered. However, a comprehensive, genome-wide, study of AS regulation for the SR proteins has never been performed.

In chapter two, we profiled the changes in AS when individual SR protein expression was reduced using RNA-Seq. We identified 561 simple AS events regulated by the eight SR proteins in *Drosophila* S2 cells. We found that multiple members of the SR protein family regulate a majority of these AS targets. In addition, we found that SR proteins co-regulate the AS of the targets in the same direction, and that there is very little anti-regulation between the SR protein family members. These discoveries extend the observation of overlap in AS targets between two SR proteins in a previous study, and would have been missed with the analysis of a single SR protein [116].

By analyzing the magnitude and direction of splicing change when the expression of two SR proteins are knocked-down together on co-regulated targets, we revealed that the contribution of individual SR proteins are additive, and the combination of SR protein regulation results in the final pattern of AS observed. The expression of two SR proteins, or any RBPs, reduced at the same time has never been profiled with respect to AS. Therefore, it would be interesting to see if this combinatorial mechanism of regulation can be extended to other splicing regulators.

Using RNA-Seq we can monitor other changes in the transcriptome, when an SR protein is depleted, in addition to AS. We found that SR proteins could regulate promoter choice, poly(A) site usage, and overall transcript levels of specific genes. SR proteins are presumably regulating these events independent of their canonical interactions with the spliceosome. It has been demonstrated that SR proteins can associate with the CTD of RNA Pol II, and that other splicing regulators can associate with the cleavage and polyadenylation machinery to influence APA [120]. Since splicing and 3' end processing can occur co-transcriptionally, it will be interesting for further studies to identify the

mechanism of SR protein regulation of these events, and potentially identify further cross-regulation between transcription and downstream RNA-processing steps.

5.3 AS regulation requires an RS-domain, and specificity is determined by the RRM-domain

SR proteins have a modular protein domain structure, consisting of an N-terminal RNA-binding domain (one or two RRM-domains) and a C-terminal domain rich in arginine/serine repeats (RS-domain). In order to regulate AS, it has been shown that SR proteins bind RNA through the RRM-domain, and recruit spliceosomal components through protein-protein interactions via the RS-domain. However, previous studies have shown that the RS-domain can contact the RNA, and that the RS-domain is dispensable for the regulation of certain splicing events [133, 137]. One major flaw of these studies is that they were performed using a small number of *in vitro* splicing substrates.

Having identified numerous endogenous AS targets of the SR proteins in chapter two, in chapter three we sought to analyze the contribution of the SR protein domains to AS regulation *in vivo*. We found that in order to regulate AS *in vivo* the SR protein XL6 had to have an RS-domain. Additionally, when we substituted the RS-domain from another SR protein, B52, we found that it could functionally substitute. This is consistent with previous observations of functional RS-domain substitution. However, when we profiled AS events that are regulated antagonistically by XL6 and B52, we found that the fusion protein with the substituted RS-domain regulated the splicing pattern in the same way as XL6. This provides evidence that the RRM-domain, not the RS-domain, is responsible for determining the magnitude and direction of splicing. In addition, the fusion protein

did not regulate targets that were only affected by B52, meaning that specificity is also determined by the RRM-domain.

These observations demonstrate the importance of the RS-domain for SR protein AS regulation, *in vivo*. It also extends the observation that the RS-domains can be functionally substituted for one another, to the *Drosophila* SR proteins. Additionally, the RRM-domain determines the specificity and pattern of AS. Therefore, most likely, the RNA-binding location on the target transcript ultimately determines the magnitude and direction of splicing, and not the RS-domain. In the future, it would be interesting to examine if this behavior extends to all the SR proteins in *Drosophila*, and SR proteins in other species. It will also be interesting to use these *in vivo* targets to see how post-translational modifications of each of the SR protein domains also plays a role in regulating AS.

5.4 Global RNA-binding landscapes of the SR proteins

The observations that SR proteins cooperate with each other to regulate splicing targets, and that the RNA-binding domain contributes to the specificity and direction of splicing change, we sought out to map the global protein-RNA interactions in chapter four. Several *in vitro* approaches to identify SR protein targets have been performed, and previous studies have mapped the global, *in vivo*, protein-RNA interactions of three mammalian SR proteins. However, there has never been a systematic profile of most members of the SR protein family, and there never has been a study examining, *in vivo*, SR protein binding with regards to the regulation of endogenous AS.

Using iCLIP-Seq, we were able to map SR protein binding with high precision, and found that SR proteins bind to most transcripts in S2 cells. This is likely since SR proteins are not only important for the regulation of AS decisions, but are required for the regulation of constitutive splicing. This is also consistent with the study of other ubiquitously expressed splicing regulators, such as the hnRNP family [112]. Despite binding to many of the same genes, SR proteins are able to bind to distinct locations on the genes, often in multiple locations. The binding to genes not regulated at the level of AS, or SR protein binding to regions far away from the splice sites, indicate the possible role of SR proteins in the regulation of other RNA-processing steps. Indeed, similar to reports for mammalian SR proteins, we observed SR protein binding to many types of ncRNAs [107]. It remains to be investigated whether SR proteins are required for the processing of many ncRNAs. An interesting possibility is that the interaction with the ncRNAs can influence SR protein regulation of splicing.

The large number of *in vivo* binding sites generated by iCLIP-Seq allowed us to derive, *in vivo*, binding motifs for the SR proteins. We found that all SR proteins bind to purine-rich sequences that resemble ESEs. Some SR proteins have, distinct, specific motifs, while many of the SR proteins binding motifs are related. In addition, the motifs derived in chapter four resemble binding motifs derived for many of the mammalian SR protein homologs, suggesting conservation of SR protein binding motifs between *Drosophila* and mammals. The functionality and strict requirement for the SR protein binding motifs in the regulation of AS remains will need to be further investigated. Also, it would be interesting to identify if any of the derived binding motifs contribute to RNA secondary structure, especially since B52 has been shown to bind a hairpin loop structure [146].

Finally, we integrated our binding data with the AS information, and were able to build splicing RNA maps. As a whole, the RNA maps revealed that SR proteins primarily were associated with exonic sequences, and were bound near the splice sites that they activated. This is consistent with the notion that SR proteins are exonic splicing enhancers, and activate splicing by recruiting spliceosomal components to the nearby splice-sites.

Taken together our global analysis of SR protein binding shows that these proteins are multifunctional by binding to diverse classes of RNAs, and can regulate both constitutive and alternative exons by binding exonic regions near the activated splice site.

5.5 Final thoughts and future directions

In this dissertation we identified, genome-wide, the endogenous targets of AS, and the binding distribution of all eight SR proteins identified in *Drosophila*. We found that SR proteins regulate AS in a combinatorial manner, and bind to exonic regions near splice sites to activate splice site selection. Initially identified as splicing regulators, we found that SR proteins bind to a diverse set of RNAs, including ncRNAs, and can also regulate promoter and poly(A) site selection. This suggests a network of interactions, and extensive coupling, of different cellular programs that control gene expression. It also uncovers, yet identified, functions for SR proteins in other RNA-processing steps. These findings have defined the precise global regulatory network for the SR proteins, and provided insights into the molecular mechanism of SR protein AS regulation, and gene expression control.

In future studies, it will be interesting to examine if this type of combinatorial regulation exists beyond the SR protein family, and can be extended to many different types of splicing regulators. Many AS events are regulated tissue specifically, or at precise developmental time points, and can contribute to disease progression if misregulated. Examination of the expression pattern of SR proteins in different tissues and developmental conditions may reveal the precise mechanism of regulation of many AS events important for development and tissue homeostasis, and could provide novel targets for the treatment of disease.

In addition to SR protein's roles in the regulation of AS, they can regulate many other processes that may be coupled to AS regulation. Although we identified targets where SR proteins affect promoter selection, polyadenylation site choice, and binding to noncoding and intronless RNAs, the precise mechanism of regulation has not yet been identified. It will be interesting to assay the interaction between SR proteins and other regulatory complexes important in other gene expression regulatory networks.

Fully understanding SR protein regulation beyond AS will allow greater understanding of the integration of gene expression regulation between multiple regulatory networks, and may identify SR proteins not only as AS regulators, but master regulators of gene expression in higher eukaryotes.

Chapter VI: References

1. Black, D.L., *Mechanisms of alternative pre-messenger RNA splicing*. Annu Rev Biochem, 2003. **72**: p. 291-336.
2. Kramer, A., *The structure and function of proteins involved in mammalian pre-mRNA splicing*. Annu Rev Biochem, 1996. **65**: p. 367-409.
3. Will, C.L. and R. Luhrmann, *Spliceosomal UsnRNP biogenesis, structure and function*. Current opinion in cell biology, 2001. **13**(3): p. 290-301.
4. Das, R., Z. Zhou, and R. Reed, *Functional association of U2 snRNP with the ATP-independent spliceosomal complex E*. Mol Cell, 2000. **5**(5): p. 779-87.
5. Matlin, A.J. and M.J. Moore, *Spliceosome assembly and composition*. Adv Exp Med Biol, 2007. **623**: p. 14-35.
6. Flicek, P., et al., *Ensembl's 10th year*. Nucleic Acids Res, 2010. **38**(Database issue): p. D557-62.
7. Hahn, M.W. and G.A. Wray, *The g-value paradox*. Evolution & development, 2002. **4**(2): p. 73-5.
8. Nilsen, T.W. and B.R. Graveley, *Expansion of the eukaryotic proteome by alternative splicing*. Nature, 2010. **463**(7280): p. 457-63.
9. Wang, E.T., et al., *Alternative isoform regulation in human tissue transcriptomes*. Nature, 2008. **456**(7221): p. 470-6.
10. Pan, Q., et al., *Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing*. Nat Genet, 2008. **40**(12): p. 1413-5.
11. Baker, B.S. and J.M. Belote, *Sex determination and dosage compensation in Drosophila melanogaster*. Annual review of genetics, 1983. **17**: p. 345-93.
12. Schutt, C. and R. Nothiger, *Structure, function and evolution of sex-determining systems in Dipteran insects*. Development, 2000. **127**(4): p. 667-77.
13. Valcarcel, J., et al., *The protein Sex-lethal antagonizes the splicing factor U2AF to regulate alternative splicing of transformer pre-mRNA*. Nature, 1993. **362**(6416): p. 171-5.

14. Boutz, P.L., et al., *A post-transcriptional regulatory switch in polypyrimidine tract-binding proteins reprograms alternative splicing in developing neurons*. Genes Dev, 2007. **21**(13): p. 1636-52.
15. Chou, M.Y., et al., *Multisite RNA binding and release of polypyrimidine tract binding protein during the regulation of c-src neural-specific splicing*. Mol Cell, 2000. **5**(6): p. 949-57.
16. Rooke, N., et al., *Roles for SR proteins and hnRNP A1 in the regulation of c-src exon NI*. Mol Cell Biol, 2003. **23**(6): p. 1874-84.
17. Krawczak, M., J. Reiss, and D.N. Cooper, *The mutational spectrum of single base-pair substitutions in mRNA splice junctions of human genes: causes and consequences*. Hum Genet, 1992. **90**(1-2): p. 41-54.
18. Cartegni, L. and A.R. Krainer, *Disruption of an SF2/ASF-dependent exonic splicing enhancer in SMN2 causes spinal muscular atrophy in the absence of SMN1*. Nat Genet, 2002. **30**(4): p. 377-84.
19. Kashima, T. and J.L. Manley, *A negative element in SMN2 exon 7 inhibits splicing in spinal muscular atrophy*. Nat Genet, 2003. **34**(4): p. 460-3.
20. Wirth, B., L. Brichta, and E. Hahnen, *Spinal muscular atrophy: from gene to therapy*. Seminars in pediatric neurology, 2006. **13**(2): p. 121-31.
21. Cartegni, L. and A.R. Krainer, *Correction of disease-associated exon skipping by synthetic exon-specific activators*. Nature structural biology, 2003. **10**(2): p. 120-5.
22. Boise, L.H., et al., *bcl-x, a bcl-2-related gene that functions as a dominant regulator of apoptotic cell death*. Cell, 1993. **74**(4): p. 597-608.
23. Takehara, T., et al., *Expression and role of Bcl-xL in human hepatocellular carcinomas*. Hepatology, 2001. **34**(1): p. 55-61.
24. Xerri, L., et al., *Predominant expression of the long isoform of Bcl-x (Bcl-xL) in human lymphomas*. British journal of haematology, 1996. **92**(4): p. 900-6.
25. Karni, R., et al., *The gene encoding the splicing factor SF2/ASF is a proto-oncogene*. Nat Struct Mol Biol, 2007. **14**(3): p. 185-93.
26. Misquitta-Ali, C.M., et al., *Global profiling and molecular characterization of alternative splicing events misregulated in lung cancer*. Mol Cell Biol, 2011. **31**(1): p. 138-50.
27. Paronetto, M.P., et al., *The RNA-binding protein Sam68 modulates the alternative splicing of Bcl-x*. The Journal of cell biology, 2007. **176**(7): p. 929-39.

28. Keren, H., G. Lev-Maor, and G. Ast, *Alternative splicing and evolution: diversification, exon definition and function*. Nature reviews. Genetics, 2010. **11**(5): p. 345-55.
29. Koren, E., G. Lev-Maor, and G. Ast, *The emergence of alternative 3' and 5' splice site exons from constitutive exons*. PLoS Comput Biol, 2007. **3**(5): p. e95.
30. Lim, L.P. and C.B. Burge, *A computational analysis of sequence features involved in recognition of short introns*. Proceedings of the National Academy of Sciences of the United States of America, 2001. **98**(20): p. 11193-8.
31. Reed, R. and T. Maniatis, *A role for exon sequences and splice-site proximity in splice-site selection*. Cell, 1986. **46**(5): p. 681-90.
32. Berget, S.M., *Exon recognition in vertebrate splicing*. J Biol Chem, 1995. **270**(6): p. 2411-4.
33. Buratti, E. and F.E. Baralle, *Influence of RNA secondary structure on the pre-mRNA splicing process*. Mol Cell Biol, 2004. **24**(24): p. 10505-14.
34. Kornblihtt, A.R., et al., *Multiple links between transcription and splicing*. Rna, 2004. **10**(10): p. 1489-98.
35. Zhong, X.Y., et al., *SR proteins in vertical integration of gene expression from transcription to RNA processing to translation*. Mol Cell, 2009. **35**(1): p. 1-10.
36. Dreyfuss, G., et al., *hnRNP proteins and the biogenesis of mRNA*. Annu Rev Biochem, 1993. **62**: p. 289-321.
37. Chen, M. and J.L. Manley, *Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches*. Nature reviews. Molecular cell biology, 2009. **10**(11): p. 741-54.
38. Manley, J.L. and A.R. Krainer, *A rational nomenclature for serine/arginine-rich protein splicing factors (SR proteins)*. Genes Dev, 2010. **24**(11): p. 1073-4.
39. Busch, A. and K.J. Hertel, *Evolution of SR protein and hnRNP splicing regulatory factors*. Wiley interdisciplinary reviews. RNA, 2012. **3**(1): p. 1-12.
40. Cáceres, J.F. and A.R. Krainer, *Functional analysis of pre-mRNA splicing factor SF2/ASF structural domains*. Embo J, 1993. **12**(12): p. 4715-26.
41. Zuo, P. and J.L. Manley, *Functional domains of the human splicing factor ASF/SF2*. Embo J, 1993. **12**(12): p. 4727-37.
42. Krainer, A.R., G.C. Conway, and D. Kozak, *The essential pre-mRNA splicing factor SF2 influences 5' splice site selection by activating proximal sites*. Cell, 1990. **62**(1): p. 35-42.

43. Long, J.C. and J.F. Cáceres, *The SR protein family of splicing factors: master regulators of gene expression*. Biochem J, 2009. **417**(1): p. 15-27.
44. Shepard, P.J. and K.J. Hertel, *The SR protein family*. Genome biology, 2009. **10**(10): p. 242.
45. Lin, S. and X.D. Fu, *SR proteins and related factors in alternative splicing*. Adv Exp Med Biol, 2007. **623**: p. 107-22.
46. Moroy, T. and F. Heyd, *The impact of alternative splicing in vivo: mouse models show the way*. Rna, 2007. **13**(8): p. 1155-71.
47. Ring, H.Z. and J.T. Lis, *The SR protein B52/SRp55 is essential for Drosophila development*. Mol Cell Biol, 1994. **14**(11): p. 7499-506.
48. Ding, J.H., et al., *Dilated cardiomyopathy caused by tissue-specific ablation of SC35 in the heart*. Embo J, 2004. **23**(4): p. 885-96.
49. Xu, X., et al., *ASF/SF2-regulated CaMKII δ alternative splicing temporally reprograms excitation-contraction coupling in cardiac muscle*. Cell, 2005. **120**(1): p. 59-72.
50. Gabut, M., et al., *The SR family proteins B52 and dASF/SF2 modulate development of the Drosophila visual system by regulating specific RNA targets*. Mol Cell Biol, 2007. **27**(8): p. 3087-97.
51. Fischer, D.C., et al., *Expression of splicing factors in human ovarian cancer*. Oncology reports, 2004. **11**(5): p. 1085-90.
52. Stickeler, E., et al., *Stage-specific changes in SR splicing factors and alternative splicing in mammary tumorigenesis*. Oncogene, 1999. **18**(24): p. 3574-82.
53. Graveley, B.R., *Sorting out the complexity of SR protein functions*. Rna, 2000. **6**(9): p. 1197-211.
54. Lam, B.J. and K.J. Hertel, *A general role for splicing enhancers in exon definition*. Rna, 2002. **8**(10): p. 1233-41.
55. Matlin, A.J., F. Clark, and C.W. Smith, *Understanding alternative splicing: towards a cellular code*. Nat Rev Mol Cell Biol, 2005. **6**(5): p. 386-98.
56. Wu, J.Y. and T. Maniatis, *Specific interactions between proteins implicated in splice site selection and regulated alternative splicing*. Cell, 1993. **75**(6): p. 1061-70.
57. Zhu, J., A. Mayeda, and A.R. Krainer, *Exon identity established through differential antagonism between exonic splicing silencer-bound hnRNP A1 and enhancer-bound SR proteins*. Mol Cell, 2001. **8**(6): p. 1351-61.

58. Liu, H.X., M. Zhang, and A.R. Krainer, *Identification of functional exonic splicing enhancer motifs recognized by individual SR proteins*. Genes Dev, 1998. **12**(13): p. 1998-2012.
59. Goren, A., et al., *Comparative analysis identifies exonic splicing regulatory sequences--The complex definition of enhancers and silencers*. Mol Cell, 2006. **22**(6): p. 769-81.
60. Graveley, B.R., K.J. Hertel, and T. Maniatis, *A systematic analysis of the factors that determine the strength of pre-mRNA splicing enhancers*. Embo J, 1998. **17**(22): p. 6747-56.
61. Lynch, K.W. and T. Maniatis, *Assembly of specific SR protein complexes on distinct regulatory elements of the Drosophila doublesex splicing enhancer*. Genes Dev, 1996. **10**(16): p. 2089-101.
62. Expert-Bezancon, A., et al., *hnRNP A1 and the SR proteins ASF/SF2 and SC35 have antagonistic functions in splicing of beta-tropomyosin exon 6B*. J Biol Chem, 2004. **279**(37): p. 38249-59.
63. Soret, J. and J. Tazi, *Phosphorylation-dependent control of the pre-mRNA splicing machinery*. Progress in molecular and subcellular biology, 2003. **31**: p. 89-126.
64. Mattaj, I.W., *RNA processing. Splicing in space*. Nature, 1994. **372**(6508): p. 727-8.
65. Colwill, K., et al., *The Clk/Sty protein kinase phosphorylates SR splicing factors and regulates their intranuclear distribution*. Embo J, 1996. **15**(2): p. 265-75.
66. Huang, Y., et al., *SR splicing factors serve as adapter proteins for TAP-dependent mRNA export*. Mol Cell, 2003. **11**(3): p. 837-43.
67. Huang, Y. and J.A. Steitz, *Splicing factors SRp20 and 9G8 promote the nucleocytoplasmic export of mRNA*. Mol Cell, 2001. **7**(4): p. 899-905.
68. Michlewski, G., J.R. Sanford, and J.F. Cáceres, *The splicing factor SF2/ASF regulates translation initiation by enhancing phosphorylation of 4E-BP1*. Mol Cell, 2008. **30**(2): p. 179-89.
69. Bedard, K.M., S. Daijogo, and B.L. Semler, *A nucleo-cytoplasmic SR protein functions in viral IRES-mediated translation initiation*. Embo J, 2007. **26**(2): p. 459-67.
70. Swartz, J.E., et al., *The shuttling SR protein 9G8 plays a role in translation of unspliced mRNA containing a constitutive transport element*. J Biol Chem, 2007. **282**(27): p. 19844-53.

71. Zhang, Z. and A.R. Krainer, *Involvement of SR proteins in mRNA surveillance*. Mol Cell, 2004. **16**(4): p. 597-607.
72. Lin, S., et al., *The splicing factor SC35 has an active role in transcriptional elongation*. Nat Struct Mol Biol, 2008. **15**(8): p. 819-26.
73. Johnson, J.M., et al., *Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays*. Science, 2003. **302**(5653): p. 2141-4.
74. Blanchette, M., et al., *Genome-wide analysis of alternative pre-mRNA splicing and RNA-binding specificities of the Drosophila hnRNP A/B family members*. Mol Cell, 2009. **33**(4): p. 438-49.
75. Ule, J., et al., *Nova regulates brain-specific splicing to shape the synapse*. Nat Genet, 2005. **37**(8): p. 844-52.
76. Agarwal, A., et al., *Comparison and calibration of transcriptome data from RNA-Seq and tiling arrays*. BMC genomics, 2010. **11**: p. 383.
77. Okoniewski, M.J. and C.J. Miller, *Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations*. BMC bioinformatics, 2006. **7**: p. 276.
78. Mortazavi, A., et al., *Mapping and quantifying mammalian transcriptomes by RNA-Seq*. Nature methods, 2008. **5**(7): p. 621-8.
79. Morin, R., et al., *Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing*. BioTechniques, 2008. **45**(1): p. 81-94.
80. Li, Z., et al., *RNA-Seq improves annotation of protein-coding genes in the cucumber genome*. BMC genomics, 2011. **12**: p. 540.
81. Li, H., J. Ruan, and R. Durbin, *Mapping short DNA sequencing reads and calling variants using mapping quality scores*. Genome Res, 2008. **18**(11): p. 1851-8.
82. Langmead, B., et al., *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome*. Genome biology, 2009. **10**(3): p. R25.
83. Brooks, A.N., et al., *Conservation of an RNA regulatory map between Drosophila and mammals*. Genome Res, 2011. **21**(2): p. 193-202.
84. Katz, Y., et al., *Analysis and design of RNA sequencing experiments for identifying isoform regulation*. Nature methods, 2010. **7**(12): p. 1009-15.
85. Trapnell, C., et al., *Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation*. Nat Biotechnol, 2010. **28**(5): p. 511-5.

86. Nagalakshmi, U., et al., *The transcriptional landscape of the yeast genome defined by RNA sequencing*. Science, 2008. **320**(5881): p. 1344-9.
87. Cheah, M.T., et al., *Control of alternative RNA splicing and gene expression by eukaryotic riboswitches*. Nature, 2007. **447**(7143): p. 497-500.
88. Kishore, S. and S. Stamm, *The snoRNA HBII-52 regulates alternative splicing of the serotonin receptor 2C*. Science, 2006. **311**(5758): p. 230-2.
89. Fairbrother, W.G., et al., *Predictive identification of exonic splicing enhancers in human genes*. Science, 2002. **297**(5583): p. 1007-13.
90. Zhang, X.H. and L.A. Chasin, *Computational definition of sequence motifs governing constitutive exon splicing*. Genes Dev, 2004. **18**(11): p. 1241-50.
91. Zhang, C., et al., *Defining the regulatory network of the tissue-specific splicing factors Fox-1 and Fox-2*. Genes Dev, 2008. **22**(18): p. 2550-63.
92. Tuerk, C. and L. Gold, *Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase*. Science, 1990. **249**(4968): p. 505-10.
93. Tacke, R., Y. Chen, and J.L. Manley, *Sequence-specific RNA binding by an SR protein requires RS domain phosphorylation: creation of an SRp40-specific splicing enhancer*. Proceedings of the National Academy of Sciences of the United States of America, 1997. **94**(4): p. 1148-53.
94. Tacke, R. and J.L. Manley, *The human splicing factors ASF/SF2 and SC35 possess distinct, functionally significant RNA binding specificities*. Embo J, 1995. **14**(14): p. 3540-51.
95. Cavalc, Y., et al., *The splicing factors 9G8 and SRp20 transactivate splicing through different and specific enhancers*. Rna, 1999. **5**(3): p. 468-83.
96. Coulter, L.R., M.A. Landree, and T.A. Cooper, *Identification of a new class of exonic splicing enhancers by in vivo selection*. Mol Cell Biol, 1997. **17**(4): p. 2143-50.
97. Niranjankumari, S., et al., *Reversible cross-linking combined with immunoprecipitation to study RNA-protein interactions in vivo*. Methods, 2002. **26**(2): p. 182-90.
98. Mili, S. and J.A. Steitz, *Evidence for reassociation of RNA-binding proteins after cell lysis: implications for the interpretation of immunoprecipitation analyses*. Rna, 2004. **10**(11): p. 1692-4.
99. Ule, J., et al., *CLIP identifies Nova-regulated RNA networks in the brain*. Science, 2003. **302**(5648): p. 1212-5.

100. Licatalosi, D.D., et al., *HITS-CLIP yields genome-wide insights into brain alternative RNA processing*. Nature, 2008. **456**(7221): p. 464-9.
101. Sanford, J.R., et al., *Splicing factor SFRS1 recognizes a functionally diverse landscape of RNA transcripts*. Genome Res, 2009. **19**(3): p. 381-94.
102. Sugimoto, Y., et al., *Analysis of CLIP and iCLIP methods for nucleotide-resolution studies of protein-RNA interactions*. Genome biology, 2012. **13**(8): p. R67.
103. Hafner, M., et al., *Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP*. Cell, 2010. **141**(1): p. 129-41.
104. Konig, J., et al., *iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution*. Nat Struct Mol Biol, 2010. **17**(7): p. 909-15.
105. Urlaub, H., K. Hartmuth, and R. Luhrmann, *A two-tracked approach to analyze RNA-protein crosslinking sites in native, nonlabeled small nuclear ribonucleoprotein particles*. Methods, 2002. **26**(2): p. 170-81.
106. Wang, Z., et al., *iCLIP predicts the dual splicing effects of TIA-RNA interactions*. PLoS biology, 2010. **8**(10): p. e1000530.
107. Anko, M.L., et al., *The RNA-binding landscapes of two SR proteins reveal unique functions and binding to diverse RNA classes*. Genome biology, 2012. **13**(3): p. R17.
108. Kanopka, A., O. Muhlemann, and G. Akusjarvi, *Inhibition by SR proteins of splicing of a regulated adenovirus pre-mRNA*. Nature, 1996. **381**(6582): p. 535-8.
109. Hui, J., et al., *Intronic CA-repeat and CA-rich elements: a new class of regulators of mammalian alternative splicing*. Embo J, 2005. **24**(11): p. 1988-98.
110. Witten, J.T. and J. Ule, *Understanding splicing regulation through RNA splicing maps*. Trends in genetics : TIG, 2011. **27**(3): p. 89-97.
111. Dredge, B.K., et al., *Nova autoregulation reveals dual functions in neuronal splicing*. Embo J, 2005. **24**(8): p. 1608-20.
112. Huelga, S.C., et al., *Integrative genome-wide analysis reveals cooperative regulation of alternative splicing by hnRNP proteins*. Cell reports, 2012. **1**(2): p. 167-178.
113. Yeo, G.W., et al., *An RNA code for the FOX2 splicing regulator revealed by mapping RNA-protein interactions in stem cells*. Nat Struct Mol Biol, 2009. **16**(2): p. 130-7.

114. Xue, Y., et al., *Genome-wide analysis of PTB-RNA interactions reveals a strategy used by the general splicing repressor to modulate exon inclusion or skipping*. Mol Cell, 2009. **36**(6): p. 996-1006.
115. Zahler, A.M., et al., *SR proteins: a conserved family of pre-mRNA splicing factors*. Genes Dev, 1992. **6**(5): p. 837-47.
116. Blanchette, M., et al., *Global analysis of positive and negative pre-mRNA splicing regulators in Drosophila*. Genes Dev, 2005. **19**(11): p. 1306-14.
117. Wang, J., Y. Takagaki, and J.L. Manley, *Targeted disruption of an essential vertebrate gene: ASF/SF2 is required for cell viability*. Genes Dev, 1996. **10**(20): p. 2588-99.
118. Wang, J. and J.L. Manley, *Overexpression of the SR proteins ASF/SF2 and SC35 influences alternative splicing in vivo in diverse ways*. Rna, 1995. **1**(3): p. 335-46.
119. Schibler, U. and F. Sierra, *Alternative promoters in developmental gene expression*. Annual review of genetics, 1987. **21**: p. 237-57.
120. Di Giammartino, D.C., K. Nishida, and J.L. Manley, *Mechanisms and consequences of alternative polyadenylation*. Mol Cell, 2011. **43**(6): p. 853-66.
121. Alt, F.W., et al., *Synthesis of secreted and membrane-bound immunoglobulin mu heavy chains is directed by mRNAs that differ at their 3' ends*. Cell, 1980. **20**(2): p. 293-301.
122. Sandberg, R., et al., *Proliferating cells express mRNAs with shortened 3' untranslated regions and fewer microRNA target sites*. Science, 2008. **320**(5883): p. 1643-7.
123. Bee, T., et al., *Nonredundant roles for Runx1 alternative promoters reflect their activity at discrete stages of developmental hematopoiesis*. Blood, 2010. **115**(15): p. 3042-50.
124. Yuryev, A., et al., *The C-terminal domain of the largest subunit of RNA polymerase II interacts with a novel set of serine/arginine-rich proteins*. Proceedings of the National Academy of Sciences of the United States of America, 1996. **93**(14): p. 6975-80.
125. Kumar, S. and A.J. Lopez, *Negative feedback regulation among SR splicing factors encoded by Rbp1 and Rbp1-like in Drosophila*. Embo J, 2005. **24**(14): p. 2646-55.
126. Hoffman, B.E. and J.T. Lis, *Pre-mRNA splicing by the essential Drosophila protein B52: tissue and target specificity*. Mol Cell Biol, 2000. **20**(1): p. 181-6.

127. Labourier, E., et al., *Antagonism between RSF1 and SR proteins for both splice-site recognition in vitro and Drosophila development*. Genes Dev, 1999. **13**(6): p. 740-53.
128. Lou, H., et al., *Regulation of alternative polyadenylation by UI snRNPs and SRp20*. Mol Cell Biol, 1998. **18**(9): p. 4977-85.
129. Champlin, D.T., et al., *Characterization of a Drosophila protein associated with boundaries of transcriptionally active chromatin*. Genes Dev, 1991. **5**(9): p. 1611-21.
130. Fu, X.D., et al., *General splicing factors SF2 and SC35 have equivalent activities in vitro, and both affect alternative 5' and 3' splice site selection*. Proceedings of the National Academy of Sciences of the United States of America, 1992. **89**(23): p. 11224-8.
131. Reed, R., *Initial splice-site recognition and pairing during pre-mRNA splicing*. Current opinion in genetics & development, 1996. **6**(2): p. 215-20.
132. Kohtz, J.D., et al., *Protein-protein interactions and 5'-splice-site recognition in mammalian mRNA precursors*. Nature, 1994. **368**(6467): p. 119-24.
133. Shen, H. and M.R. Green, *RS domains contact splicing signals and promote splicing by a common mechanism in yeast through humans*. Genes Dev, 2006. **20**(13): p. 1755-65.
134. Shen, H. and M.R. Green, *A pathway of sequential arginine-serine-rich domain-splicing signal interactions during mammalian spliceosome assembly*. Mol Cell, 2004. **16**(3): p. 363-73.
135. Cho, S., et al., *Interaction between the RNA binding domains of Ser-Arg splicing factor 1 and UI-70K snRNP protein determines early spliceosome assembly*. Proceedings of the National Academy of Sciences of the United States of America, 2011. **108**(20): p. 8233-8.
136. Xiao, S.H. and J.L. Manley, *Phosphorylation of the ASF/SF2 RS domain affects both protein-protein and protein-RNA interactions and is necessary for splicing*. Genes Dev, 1997. **11**(3): p. 334-44.
137. Zhu, J. and A.R. Krainer, *Pre-mRNA splicing in the absence of an SR protein RS domain*. Genes Dev, 2000. **14**(24): p. 3166-78.
138. Wang, J., S.H. Xiao, and J.L. Manley, *Genetic analysis of the SR protein ASF/SF2: interchangeability of RS domains and negative control of splicing*. Genes Dev, 1998. **12**(14): p. 2222-33.

139. Mermoud, J.E., P.T. Cohen, and A.I. Lamond, *Regulation of mammalian spliceosome assembly by a protein phosphorylation mechanism*. *Embo J*, 1994. **13**(23): p. 5679-88.
140. Moore, M.J., *From birth to death: the complex lives of eukaryotic mRNAs*. *Science*, 2005. **309**(5740): p. 1514-8.
141. Polymenidou, M., et al., *Long pre-mRNA depletion and RNA missplicing contribute to neuronal vulnerability from loss of TDP-43*. *Nature neuroscience*, 2011. **14**(4): p. 459-68.
142. Bailey, T.L., *DREME: motif discovery in transcription factor ChIP-seq data*. *Bioinformatics*, 2011. **27**(12): p. 1653-9.
143. Tripathi, V., et al., *The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation*. *Mol Cell*, 2010. **39**(6): p. 925-38.
144. Wu, H., et al., *A splicing-independent function of SF2/ASF in microRNA processing*. *Mol Cell*, 2010. **38**(1): p. 67-77.
145. Sanford, J.R., et al., *Identification of nuclear and cytoplasmic mRNA targets for the shuttling protein SF2/ASF*. *PloS one*, 2008. **3**(10): p. e3369.
146. Shi, H., B.E. Hoffman, and J.T. Lis, *A specific RNA hairpin loop structure binds the RNA recognition motifs of the Drosophila SR protein B52*. *Mol Cell Biol*, 1997. **17**(5): p. 2649-57.
147. Ghigna, C., et al., *Cell motility is controlled by SF2/ASF through alternative splicing of the Ron protooncogene*. *Mol Cell*, 2005. **20**(6): p. 881-90.
148. Bakkour, N., et al., *Small-molecule inhibition of HIV pre-mRNA splicing as a novel antiretroviral therapy to overcome drug resistance*. *PLoS pathogens*, 2007. **3**(10): p. 1530-9.
149. Neugebauer, K.M., et al., *SR proteins are autoantigens in patients with systemic lupus erythematosus. Importance of phosphoepitopes*. *Arthritis and rheumatism*, 2000. **43**(8): p. 1768-78.
150. Buratti, E., et al., *SR protein-mediated inhibition of CFTR exon 9 inclusion: molecular characterization of the intronic splicing silencer*. *Nucleic Acids Res*, 2007. **35**(13): p. 4359-68.